NAIST-IS-DD0361218

Doctoral Dissertation

Rapid Unsupervised Speaker Adaptation Based on Sufficient Statistics of Hidden Markov Models

Randy Gomez

September 29, 2006

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Randy Gomez

Thesis Committee:

Prof. Kiyohiro Shikano	(Supervisor)
Prof. Masatsugu Kidode	(Co-supervisor)
Assoc. Prof. Hiroshi Saruwatari	(member)

Rapid Unsupervised Speaker Adaptation Based on Sufficient Statistics of Hidden Markov Models[*]

Randy Gomez

Abstract

In realizing a speech recognition system robust to variation of speakers, an efficient adaptation algorithm is needed. Most adaptation techniques require many adaptation data to carry out an adaptation task. Adaptation data are often collected from the actual speaker itself in several utterances. With the time needed to gather and transcribe the adaptation utterances, together with the actual execution time of the adaptation algorithm, real-time speech recognition is difficult to realize.

We propose a novel approach in solving the problem that hinders practical implementation of speaker adaptation by using only a single untranscribed utterance from the user. This unsupervised speaker adaptation approach can execute in few seconds with a significant improvement in recognition performance as compared to data-greedy and time-exhausting adaptation schemes. This thesis, details the science behind the development and implementation of the rapid unsupervised speaker adaptation based on Hidden Markov Models-Sufficient Statistics (HMM-Sufficient Statistics).

In this approach, we process in advance the training database into HMM-Sufficient Statistics Sufficient. During the actual adaptation (online), the process starts with the N-best speaker selection which is acoustically close to the user's utterance. The HMM-Sufficient Statistics of the N-best speakers are selected

[*] Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0361218, September 29, 2006.

as adaptation data. In view of the fact that HMM-Sufficient Statistics are precomputed offline, considerable amount of computation time needed for processing is saved and re-allocated efficiently to using good-performance but computationally expensive adaptation platforms. The end result, a rapid adaptation system with good recognition performance. Experiments using Vocal Tract Length Normalization (VTLN), Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) were performed. Moreover we tested for robustness under noisy environment conditions such as office, car, crowd and booth noise in several signal-to-noise ratios (SNRs).

In this thesis we successfully designed a rapid unsupervised speaker adaptation that requires only a single arbitrary utterance without transcriptions and execute in 7 sec of adaptation time. The proposed method is suitable for speech recognition applications where adaptation data is scarce and execution time is critical. Furthermore, we have fully integrated the proposed approach in a real application using a dialogue system, where the adaptation technique is integrated and interacts freely with the recognizer and several processes in the system in a real environment condition.

Keywords:

Contents

1. Introduction **1**

1.1 The Speech Recognition System 2

1.2 Challenges in Speech Recognition 4

1.3 Thesis Overview 4

2. Speech Recognition In Adverse Conditions **6**

2.1 Problems In Speech Recognition 6

2.2 Noisy Environments 6

2.3 Speaker Variation 8

2.4 Data and Time Constraints 10

2.5 Approaches to the Problem 10

2.5.1 Speech Enhancement 10

2.5.2 Speaker Normalization and Adaptation 18

2.5.3 Minimum data and Rapid Approach 19

2.6 A Closer Look On Mismatch 19

2.7 The Proposed Approach 21

2.7.1 Under Real Environment Condition 21

2.7.2 Robustness in Speaker Variation 22

2.7.3 Rapid Adaptation Using Single Adaptation Utterance . . . 24

2.8 Summary 24

3. Statistics Theory in Speech Recognition and the Proposed Rapid Unsupervised Speaker Adaptation **26**

3.1 Bayes' Decision Theory 26

3.2 Bayesian Estimation 30

3.3 Maximum Likelihood Estimation 32

3.3.1 Expectation Maximization 34

3.4 Maximum Likelihood or Bayes' Method 41

3.5 Sufficient Statistics Background 43

3.6 Applications of Sufficient Statistics 45

3.6.1 Data Reduction 46

3.6.2 Sufficient Statistics as a Better Estimator 47

3.7 Summary 48

4. Hidden Markov Models-Sufficient Statistics and N-best Speakers Selection 49

4.1 The Expectation-step 49

4.1.1 The Forward and Backward Procedure 49

4.1.2 HMM-Sufficient Statistics Parameters 53

4.2 N-best Speakers' HMM-Sufficient Statistics as Adaptation Data . 54

4.3 Summary 57

5. Rapid Baum-Welch Unsupervised Speaker Adaptation based on N-best Speakers' HMM-Sufficient Statistics 60

5.1 Single-template HMM-Sufficient Statistics 62

5.1.1 Results 63

5.1.2 Limitations of the Single-template HMM-Sufficient Statistics Adaptation 63

5.2 Multi-Template HMM-Sufficient Statistics 65

5.2.1 Results 66

5.2.2 Limitations of the Multi-template HMM-Sufficient Statistics Adaptation 67

5.3 Weighting of the HMM-Sufficient Statistics 68

5.3.1 Results 72

5.3.2 Limitations of Weighting 72

5.4 Linear Interpolation of the Global HMM-Sufficient Statistics and Clustering of Speakers 74

5.4.1 Results 81

5.4.2 Limitations of Interpolation and Clustering 81

5.5 Summary 82

6. Rapid MLLR Unsupervised Speaker Adaptation Based on N-best Speakers' HMM-Sufficient Statistics 83

6.1 Regression Class Trees 84

6.2 Mean Adaptation 85

6.3 Variance and Covariance Adaptation 86

6.4 MLLR Adaptation Parameters Using HMM-Sufficient Statistics . 87
6.5 MLLR with Multi-template HMM-Sufficient Statistics 87
6.5.1 Results 88
6.6 Linear Interpolation of the Global HMM-Sufficient Statistics in MLLR 89
6.6.1 Results 90
6.7 Summary 90

7. Experimental Results and Discussion 92
7.1 Experimental Setup 92
7.2 Basic Results using HMM-Sufficient Statistics Adaptation 95
7.3 Adaptation Time of Conventional MLLR 96
7.4 Detailed Results on Speaker Clustering and Linear Interpolation . 97
7.5 Detailed Comparison Between the Rapid Baum-Welch and Rapid MLLR based Adaptation 99
7.6 Further Investigating the HMM-Sufficient Statistics 101
7.7 Comparisons with VTLN, MAP and Conventional MLLR 102
7.8 Summary 107

8. Evaluation in Real Environment Conditions 108
8.1 Practical Implementation 108
8.2 Redesigning The Rapid Adaptation System 109
8.3 Performance of the Modified System 113
8.4 Implementation With The On-site Dialogue System 113
8.5 Summary 115

9. Conclusion 118

Acknowledgements 121

References 122

Appendix 127

A. List of Abbreviations 127

B. Expressions Used in Weighting the Sufficient Statistics 127

Publications 129

A. Journals 129

B. International Conferences 129

C. Meetings 129

List of Figures

1 On-site speech dialogue system "Takemaru-kun". 2

2 Block diagram of a basic speech recognition system. 3

3 Speech recognition in adverse conditions. 3

4 Time domain (top) and spectogram (bottom) of a clean utterance. 7

5 Spectogram of car noise (top) and the corrupted utterance (bottom). 7

6 Spectogram of crowd noise (top) and the corrupted utterance (bottom). 8

7 Spectogram of different genders and age groups. 9

8 A time-line block diagram of computation needed for a robust speech recognition system. 11

9 Plot of Word Accuracy, NRR, and MelCD in 25 dB office noise with varying α. 12

10 Plot of Word Accuracy, NRR, and MelCD in 10 dB office noise with varying α. 12

11 Plot of Word Accuracy, NRR, and MelCD in 0 dB office noise with varying α. 13

12 Block diagram of the time-domain Wiener filtering implementation. 14

13 Original noisy signal and corresponding denoised signal using Wiener filtering. 16

14 Basic block diagram of the ETSI front-end noise robustness implementation. 16

15 Plot of the MFCC of a speech segment when using ETSI and with no processing at all. 17

16 Visualization of a single-template model. 18

17 Discrepancies in the spectrum of the Japanese vowel "i" processed by SS and ETSI front-end. 20

18 Discrepancies in the spectrum of the Japanese vowel "o" processed by SS and ETSI front-end. 20

19 Mismatch in the MFCC values of a single utterance when using different processing. 21

20 Using multi-template selection and model adaptation for robustness in speaker variation. 22

21 Overview of the proposed recognition approach using only a single arbitrary utterance for adaptation. 23
22 Block diagram of basic Bayesian decision rule. 29
23 An example of an easy pattern classification problem where two classes can be linearly separated. 30
24 Comparison between Bayes' and LS when classifying two classes whose boundary cannot be separated linearly. 31
25 The Bayesian estimation method. 31
26 Overview of the E-M algorithm maximizing the objective function with a lower bound. 34
27 Basic illustration of the iterative E-M Algorithm used to find and maximize the bound. 35
28 The lower bound estimate that touches the likelihood function with iteration 1 using a good initial estimate. 36
29 The lower bound estimate that touches the likelihood function with iteration 2 using a good initial estimate. 36
30 The lower bound estimate that touches the likelihood function with iteration 3 using a good initial estimate. 37
31 The lower bound estimate that touches the likelihood function with iteration 4 using a good initial estimate. 37
32 The lower bound estimate that touches the likelihood function with iteration 1 using a random initial estimate. 38
33 The lower bound estimate that touches the likelihood function with iteration 2 using a random initial estimate. 39
34 The lower bound estimate that touches the likelihood function with iteration 3 using a random initial estimate. 39
35 The lower bound estimate that touches the likelihood function with iteration 4 using a random initial estimate. 40
36 The lower bound estimate that touches the likelihood function with iteration 5 using a random initial estimate. 40
37 The lower bound estimate that touches the likelihood function with iteration 6 using a random initial estimate. 41

38 Boundary created by Expectation-Maximization using a single Gaussian mixture. 42
39 Boundary created by Expectation-Maximization using two Gaussian mixtures. 42
40 Boundary created by Expectation-Maximization using three Gaussian mixtures. 43
41 The differences between keeping the training samples and keeping only the Sufficient Statistics. 46
42 Initialization of the forward probabilities 50
43 Computation of the forward probabilities by induction 50
44 The last phase of Forward algorithm which result to the terminal forward probabilities 51
45 The Backward procedure showing the initialization phase, backward recursion and the finalization phase. 52
46 Selecting adaptation data from the training database that are acoustically close to the user using a single utterance 54
47 Creating individual-speaker GMMs from the training database . . 55
48 GMM speaker selection using the noisy utterance. 56
49 Optimizing recognition performance using N-best speakers selection (adult and senior male database). 58
50 Optimizing recognition performance using N-best speakers selection (adult and senior female database). 59
51 Block diagram of the conventional HMM-Sufficient Statistics adaptation. 62
52 Recognition performance using single-template HMM-Sufficient Statistics adaptation with one iteration of Baum-Welch reestimation. . 63
53 Block diagram of the multi-template HMM-Sufficient Statistics adaptation. 64
54 Template selection using multi-template models gives better estimate than using single-template model. 66
55 Recognition performance using multi-template HMM-Sufficient Statistics adaptation with one iteration of Baum-Welch reestimation. . 67
56 Relationship between N-best and recognition performance. 68

57 Block diagram of weighting HMM-Sufficient Statistics prior to adaptation. 69
58 Effects of weighting the individual HMM-Sufficient Statistics using a linear weight. 70
59 Effects of weighting the individual HMM-Sufficient Statistics derived from the likelihood of the GMMs given the test utterance. . 71
60 Recognition performance with the weighted HMM-Sufficient Statistics adaptation based on one-iteration of Baum-Welch. 73
61 Illustration of the global Sufficient Statistics as the sum of all the speakers' Sufficient Statistics. 73
62 Block diagram of HMM-Sufficient Statistics adaptation with linear interpolation using individual speakers' Sufficient Statistics. . . . 74
63 Contour plot of the weighting factor for interpolation. 76
64 Creating clustered-speaker GMMs from the training database . . 77
65 HMM-Sufficient Statistics adaptation with linear interpolation using clustered speakers' Sufficient Statistics. 78
66 Clustering of speakers using Minimax. 79
67 Reduction of adaptation time when using linear interpolation of the global HMM-Sufficient Statistics. 79
68 Recognition performance of a reduced N-best without interpolation of HMM-Sufficient Statistics. 80
69 Mixture components of Gaussians from the model set that are acoustically close which are grouped together in the regression tree. 84
70 Block diagram of the online MLLR adaptation using HMM-Sufficient Statistics. 88
71 Recognition performance of the rapid MLLR adaptation based on HMM-Sufficient Statistics. 89
72 Reduction of adaptation time when using linear interpolation of the global HMM-Sufficient Statistics in the online MLLR adaptation. 91
73 Block diagram of the overall system implementation. 93
74 Overall recognition performance for 25 dB office noise environment using all platforms of HMM-Sufficient Statistics adaptation. . . . 93

75 Adaptation time of the conventional MLLR and the proposed MLLR adaptation based on HMM-Sufficient Statistics. 97

76 Recognition Performance of the Baum-Welch based linear interpolation approach. 98

77 Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Adult Male). 99

78 Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Adult Female). 100

79 Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Senior Male). 100

80 Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Senior Female). 101

81 Effects of linear interpolation of the global HMM-Sufficient Statistics. 102

82 unupdated models caused by insufficiency of adaptation data in Baum-Welch reestimation. 103

83 Block diagram of the supervised VTLN adaptation in finding for the optimum α. 104

84 Average values of α used in VTLN adaptation. 105

85 Recognition performance with various adaptation techniques. 105

86 Performance comparison of HMM-Sufficient Statistics vs. VTLN, MAP, and MLLR. 106

87 Treating individual system independently to hasten up testing during experimental phase. 109

88 Multiple processes working interdependently in an integrated system. 110

89 Testing set-up of the rapid unsupervised speaker adaptation (experimental approach). 110

90 Testing set-up of the rapid unsupervised speaker adaptation (practical implementation). 111

91 Implementation of the modified adaptation algorithm to accommodate practical issues. 112

92 Recognition performance of the practical implementation of the rapid unsupervised speaker adaptation based on HMM-Sufficient Statistics. 114

93 Actual system integration of the adaptation algorithm in the dialogue system. 116
94 Reducing execution time through effective use of memory in minimizing disk access time in loading and saving parameters. 119

List of Tables

1 System specifications 94
2 Set-up of Adult and Senior JNAS Database for HMM-Sufficient Statistics 94
3 HMM-Sufficient Statistics Info 94
4 Word Accuracy of various SNR Statistics (Single-temp: $S_{optimal} = 40$ / Multi-temp: $S = 40$ / Multi-temp with interpolation $S = 25$ /Multi-temp with interpolation, Clustered-Speakers $S = 20$ / MLLR-based 128-class. 95

1. Introduction

Automatic Speech Recognition (ASR) can be viewed as an advance form of pattern classification. In a basic pattern classification problem, one is tasked of segregating different classes of objects. Basically, in the same manner the speech utterance is broken down into sequence of sounds that vary acoustically, and the recognizer processes each of these to produce a hypothesis that correspond to the actual sound itself. Brought by the advent of fast computers, the technique of doing speech recognition has evolved rapidly. With improved computer power and disk storage, the speech recognition technology effectively combines different sources of knowledge to further increase performance and reliability. This means that in addition to the unique acoustical information the speech utterance can offer, speech recognition technology accesses and combine some other relevant sources like linguistic information in constructing the sentence in a form of semantics and pragmatics. Perhaps, context is added to validate the coherence of the recognized word with the task domain of the recognizer. There are so many sources of knowledge that is incorporated to the speech recognition system. In effect, the performance of the ASR increases significantly as the system becomes more intelligent, discriminative, and a more advanced pattern classification technique.

The availability of free softwares [1] [2], speech database and tools, revolutionalize the way this technology is applied. The once-upon-a-time fiction movies showing robots communicating to human beings is now a reality. Nowadays, whether in space flights, in health care, or perhaps in a word processor software, speech recognition is more likely being used. An example of a human-machine interface is the dialogue system in Figure 1 where speech recognition is used in answering queries and relevant information of a particular task [3]. This system is an example of a ASR that incorporates the knowledge of different sources to accomplish certain speech recognition task and to respond accordingly.

Although speech recognition has improved for the past years. Still, this field has a lot of challenging problems in attaining a very reliable result when used in real environment conditions. Efforts are being made in making it more robust and practical for all of its wide-range applications.

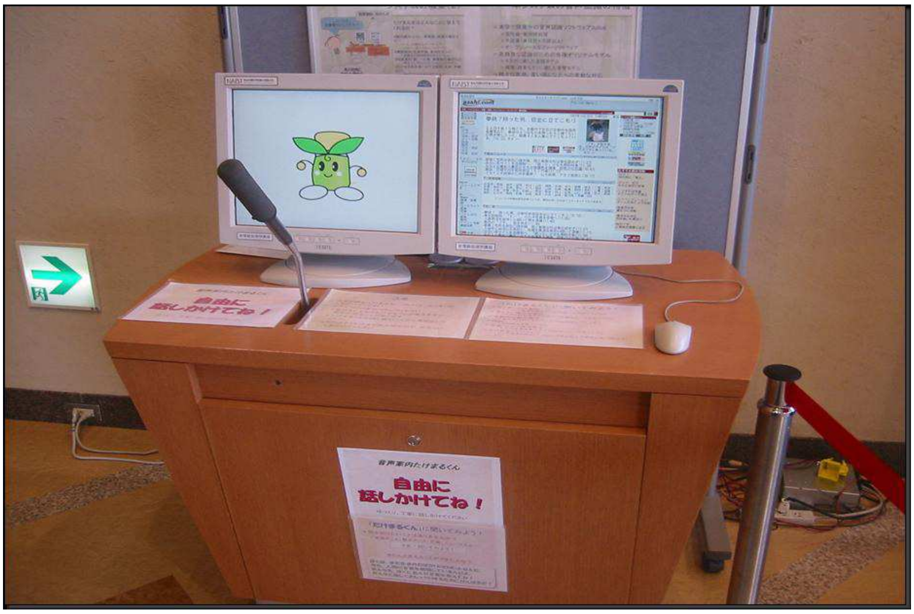


Figure 1. On-site speech dialogue system “Takemaru-kun”.

1.1 The Speech Recognition System

Techniques in speech recognition has evolved with time. These techniques include template-based system, where Dynamic Programming (DP) is used in matching an utterance with the pre-stored utterance template. Also, we have the use of *Neural Networks* often referred as the connectionist approach. The most recent and widely used technique is the probabilistic approach like the HMMs. In Figure 2 the basic speech recognition system is shown. First, the speech utterance is processed to a suitable format. Then, with the aid of the Acoustic Model and the Language Model, the recognizer establishes statistical inference which results to a hypothesis of the recognized utterance.

In general, we use a probabilistic model in which a certain $\boldsymbol{W}$ word sequence generates an acoustic observation $\boldsymbol{Y}$. The *joint probability* $P(\boldsymbol{W},\boldsymbol{Y})$ leads to Bayes’ formula:

$$P(\boldsymbol{W}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{Y})}, \tag{1}$$

where $P(\boldsymbol{Y}|\boldsymbol{W})$ is the conditional probability of the observed data given the word

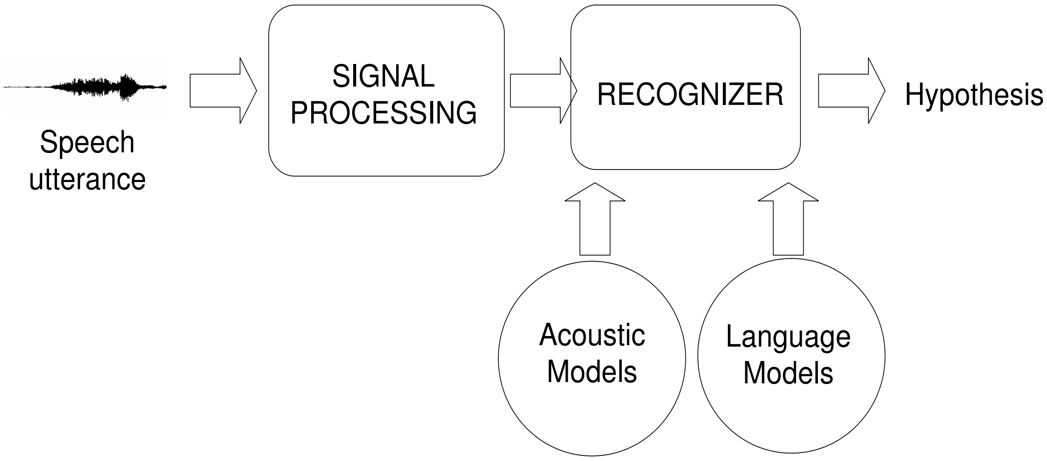


Figure 2. Block diagram of a basic speech recognition system.

Figure 3. Speech recognition in adverse conditions.

string W and generally referred to as the *acoustic model*. $P(\boldsymbol{W})$ is the a priori probability of the word sequence $\boldsymbol{W}$. Since this probability is associated with the sequence of words, it is referred to as the *language model*. The objective of the recognizer is to find

$$\widehat{\boldsymbol{W}} = arg \underset{\boldsymbol{W}}{max} P(\boldsymbol{Y}|\boldsymbol{W})P(\boldsymbol{W}). \tag{2}$$

Equation 2 is known as the MAP decoding rule which is a function of both the acoustic model and the language model. In this thesis, we deal only with the acoustic model part.

1.2 Challenges in Speech Recognition

The current speech recognition system is very sensitive to variations in the acoustic environment such as noise and non-linear channel distortions. Moreover, variation of the actual speakers' speech characteristics contribute significantly to the poor performance of the system. A scenario depicting speech recognition working in adverse conditions such as in noisy environment and with wider spectrum of users is shown in Figure 3. In short, real environment speech recognition needs more work. There are many techniques used in addressing the problems in the recognizer. However, employing additional algorithms to aid the recognizer would mean additional overhead such as computation time, and in case of adaptation techniques, adaptation data is needed to be collected and transcribed. These overheads preclude the practical aspect of speech recognition. It is imperative that in the course of formulating techniques in minimizing the effects of these problems, practical issues should be addressed.

1.3 Thesis Overview

This thesis is organized as follows, Chapter 2 presents a literature review, where the problems of speech recognition are discussed including the efforts to minimize these problems. In Chapter 3, background in statistics such as, parameter estimation is introduced which is very useful in understanding the logical progression from Maximum Likelihood Estimate, towards the basic concept of Sufficient Statistics. Theoretical background is presented together with applications in pattern recognition in general. Moreover, as the basic theoretical framework is discussed, the basic statistical concept gradually expands its application in the context of HMM. In Chapter 4, we explain the HMM-Sufficient Statistics. A detail explanation and theoretical concept in the application to adaptation using HMMs is presented, which highlights the significance and effectiveness in realizing a practical unsupervised speaker adaptation in speech recognition using only a single arbitrary adaptation utterance. Chapter 5 focuses on rapid unsupervised adaptation based on Baum-Welch reestimation, where the HMM-Sufficient Statistics is used as adaptation data. In this chapter, we discuss the basic implementation using a single template model followed by the multi-template model approach. System modifications are shown with the implementation of HMM-Sufficient Statistics weighting, global HMM-Sufficient Statistics interpolation and clustering of speakers. Each of the presented techniques are accompanied with result and discussion. Issues in recognition performance and adaptation time in the context of single iteration of Baum-Welch is the main topic in this chapter. A very different adaptation approach based on Maximum Likelihood Linear Regression (MLLR) [4] is explained in Chapter 6 when applying the concept of Sufficient Statistics. We discuss the system's design to accommodate N-best speakers' Sufficient Statistics and improve recognition performance with a very short adaptation time using MLLR. This chapter is focused on the modification of a powerful adaptation technique that takes many utterances for adaptation into a rapid adaptation technique using only a single utterance. The experimental set-up and the summary of the results in recognition performance of the proposed adaptation schemes are presented in Chapter 7. More evaluation results using current approaches like VTLN, MAP, and conventional MLLR are compared with the proposed algorithm. In Chapter 8, we show the implementation of the proposed rapid adaptation technique being integrated in an actual dialogue system. In this chapter, the actual performance of the proposed rapid unsupervised speaker adaptation technique is tested in a real environment application. Finally we conclude this thesis and describe our future work in Chapter 9.

2. Speech Recognition In Adverse Conditions

2.1 Problems In Speech Recognition

In a controlled environment condition, we assume that the intrinsic characteristics of the speech signal as it travels through the medium and into the system until its conversion to digitized form is preserved [5]. In this situation, the recognizer will be working normally as it is designed. However, this is not really the case in a real environment situation. In the physical environment, the acoustical property of the speech signal is often affected by distortion caused by additive noise which superimposes to the speech signal, and convolutive noise as well which is caused by channel distortion. Aside from these two events that happen along the medium, there are also issues that are critical in speech recognition particularly the variation of the acoustical properties of the speech inherent to the individual speaker.

2.2 Noisy Environments

Some basic speech enhancement techniques employed in speech recognition systems used to minimize the effect of noisy environment conditions are discussed in this section. For simplicity purposes, we will deal only with the additive noise.

The recognizer can achieve as much as 94.0% of recognition performance in Word Accuracy (WA) in clean environment condition using a Speaker-Independent model (SI) without any adaptation even for a 20K dictation task. However, when recognition is done in noisy environment conditions such as office, crowd, booth, and car noise, the recognition performance degrades drastically as a function of the Signal-to-Noise ratio (SNR). Thus it is important to employ speech enhancement techniques in order to minimize its effect to the recognizer. Different types of noise have different degradation impact to the speech signal. White noise tend to be easier to denoise as compared to colored noise. First, we show how the spectograms differ for a clean utterance and the noise-corrupted utterance. In Figure 4, the time domain (top) and spectogram (bottom) of a clean utterance is shown. In Figure 5 on the other hand, the spectogram of car noise (top) and spectogram of the utterance corrupted by car noise (bottom) is given. If both

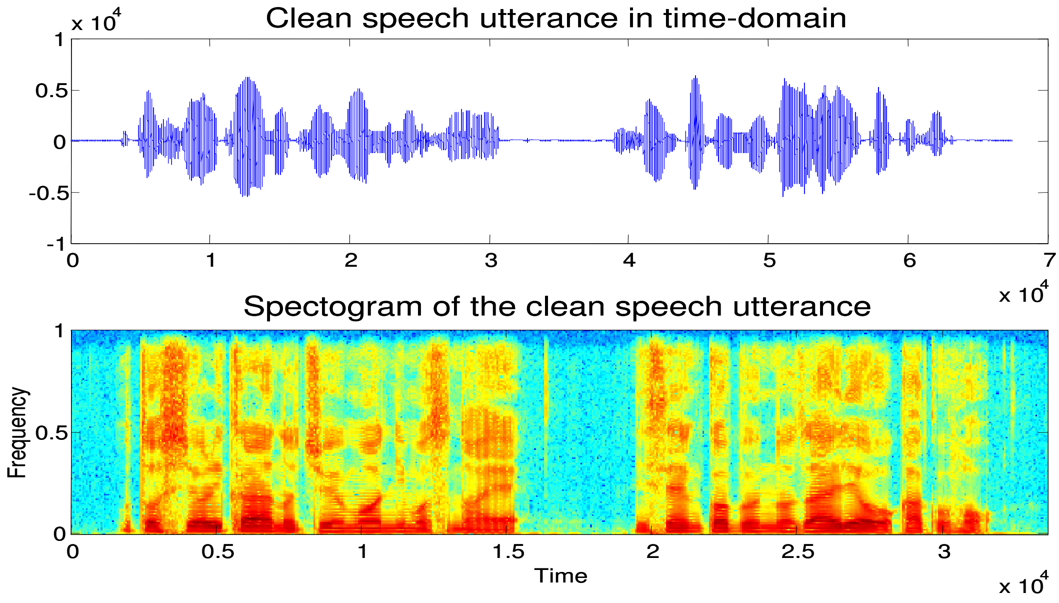


Figure 4. Time domain (top) and spectogram (bottom) of a clean utterance.

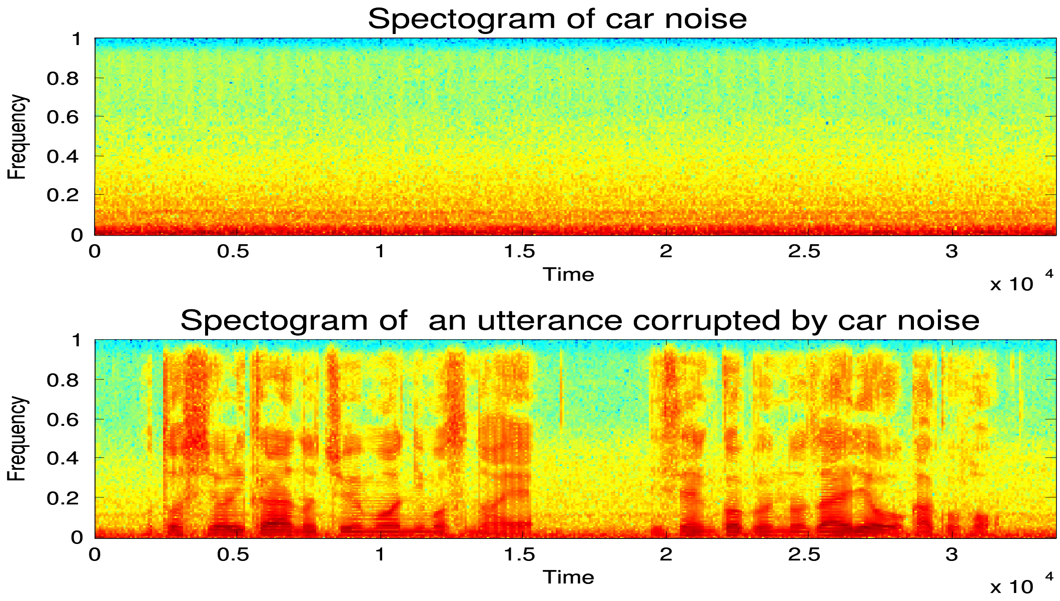


Figure 5. Spectogram of car noise (top) and the corrupted utterance (bottom).

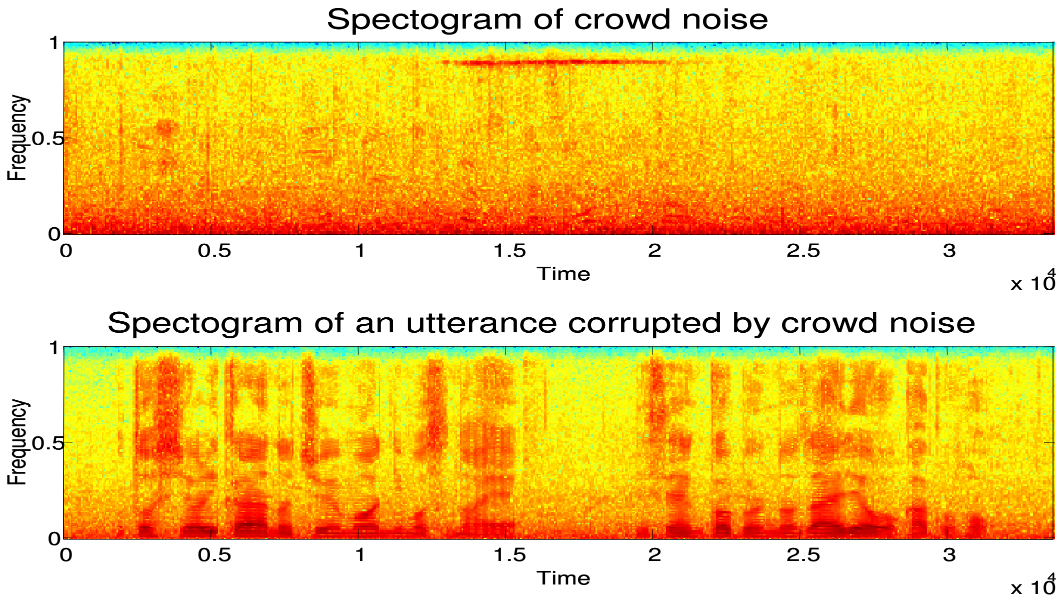


Figure 6. Spectogram of crowd noise (top) and the corrupted utterance (bottom).

Figures 4 and 5 are compared, the visible changes of the spectogram (bottom) between clean utterance and noise-corrupted utterances are apparent. Next, we show that different types of noise have different effects. Consider Figure 6 (top) where we use a different type of noise such as crowd noise. This type of noise has a speech-like nature and if we compare it to the spectogram of the car noise in Figure 5 (top), we can see that its energy is widely distributed almost over the whole frequency spectrum while the latter is just concentrated in the lower frequency part. This is one of the reasons why crowd noise is difficult to address in speech recognition, since it corrupts almost all parts of the speech in the frequency spectrum compared to some other noise, like the car noise which corrupts only the low frequency part as illustrated in Figure 5 (bottom), and Figure 6 (bottom).

2.3 Speaker Variation

In practical application, a wide variety of users are expected to use the system. Mismatch due to different age-groups and genders causes a problem of speaker

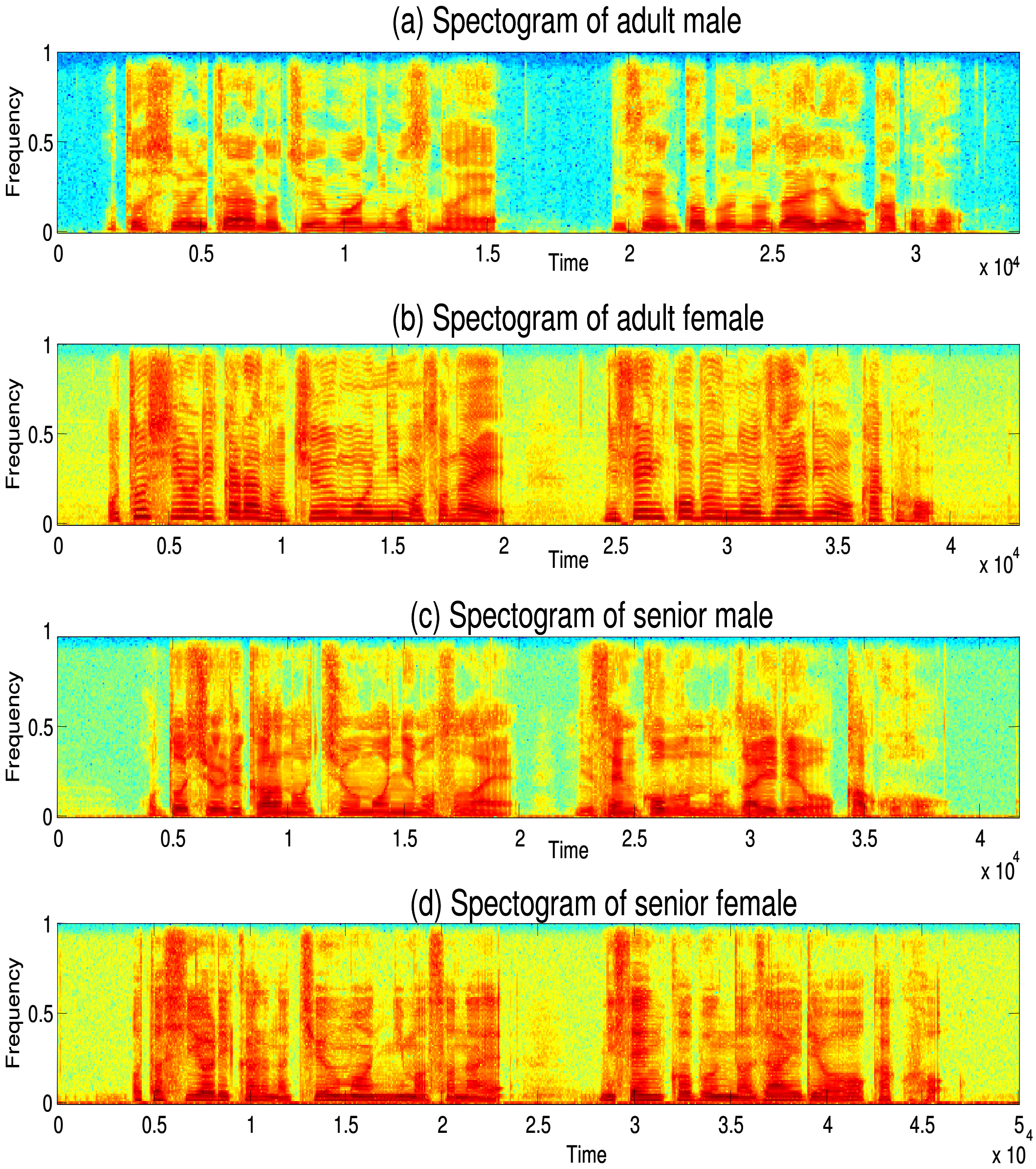


Figure 7. Spectogram of different genders and age groups.

variability which degrades the performance of the recognizer [6]. The degree of degradation depends primarily on the acoustical mismatch of the user's speech and the model. Since model-based systems are sensitive to mismatch, there is a need to employ adaptation techniques to minimize if not totally eradicate this problem. To show the variabilities of the speech utterances, Figure 7 illustrates the variations of the spectograms of the an utterance spoken by (a) adult male, (b) adult female, (c) senior male, and (d) senior female.

2.4 Data and Time Constraints

In practical applications, the challenge does not end in designing adaptation technique robust to mismatch and noisy environment, but most of all in rendering this technique practical. The algorithm used should be effective even when using minimum amount of adaptation data, and carry-out the adaptation task in a short period of time in a range of few seconds and not in several minutes. Well known adaptation techniques such as MLLR require many adaptation utterances from the test speaker himself in order to achieve a considerable improvement in recognition performance. With the time needed to gather and transcribe these adaptation utterances, together with the time to execute adaptation, real-time speech recognition is difficult to realize. To illustrate how difficult it is to design a practical system as far as data collection and adaptation time is concerned, Figure 8 shows an estimate time-line of the whole process starting from the gathering of adaptation utterances, transcribing, and parameterizing prior to the actual adaptation. If one has to roughly sum all of the accumulated time in each phase, execution time is expected to be at least in several minutes.

2.5 Approaches to the Problem

2.5.1 Speech Enhancement

One of the classical approaches in denoising speech, is the *Spectral Subtraction* (SS). Owing to its simplicity to implement [7], it has been applied in robust speech recognition under noisy environment, and is given by Equation 3

$$|S(k)|^2 = |Y(k)|^2 - \alpha|D(k)|^2, \tag{3}$$

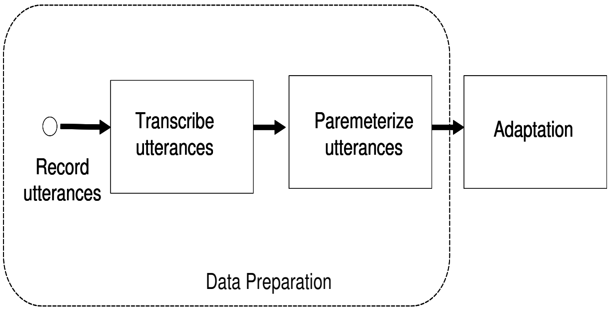


Figure 8. A time-line block diagram of computation needed for a robust speech recognition system.

where $|S(k)|$ is the denoised spectrum, $|Y(k)|$ is the noisy spectrum and |D(k)| is the noise-only spectrum. α is the oversubtraction parameter that dictates the extent of noise suppression. The oversubtraction parameter can be computed in various ways and one of these is the multi spectral approach [8]. Also, there is a method based on the human auditory system [9]. A basic expression of the oversubtraction parameter is given in Equation 4 which is a function of SNR [10]

$$\alpha = \alpha_0 - \frac{3}{2}SNR \qquad -5 \leqq SNR \leqq 25, \tag{4}$$

where α_0 is a constant.

Denoising the corrupted speech signal relatively increases the SNR but in effect, it also introduces some distortion. By using SS as a platform of denoising, we will analyze its effect as far as recognition performance is concerned in a model-based speech recognition system. Consequently, we will measure the improvement in SNR given by

$$NRR = SNR_{new} - SNR_{old} \tag{5}$$

where NRR [11] is the Noise Reduction Rate, SNR_{old} and SNR_{new} are the SNR before and after SS respectively. Also, we will consider distortion in terms of MelCD given as

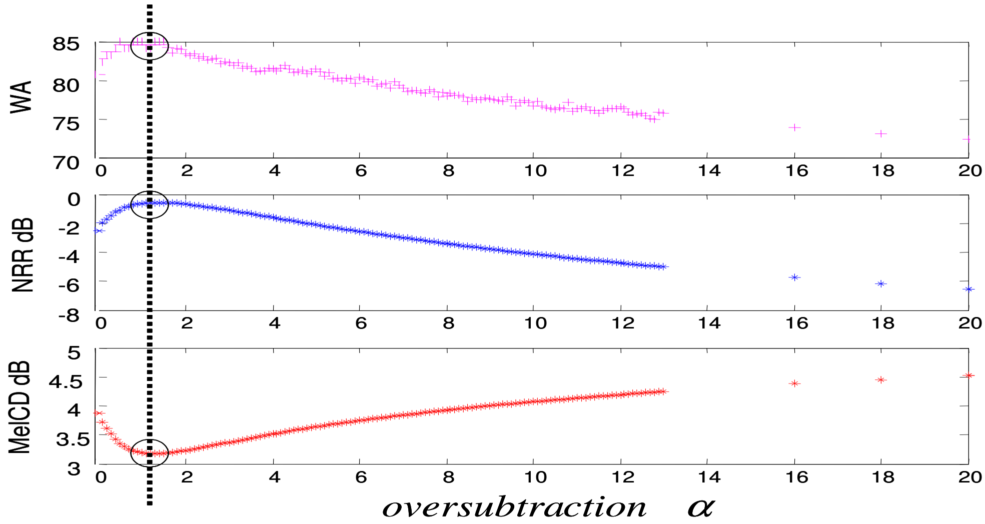


Figure 9. Plot of Word Accuracy, NRR, and MelCD in 25 dB office noise with varying α.

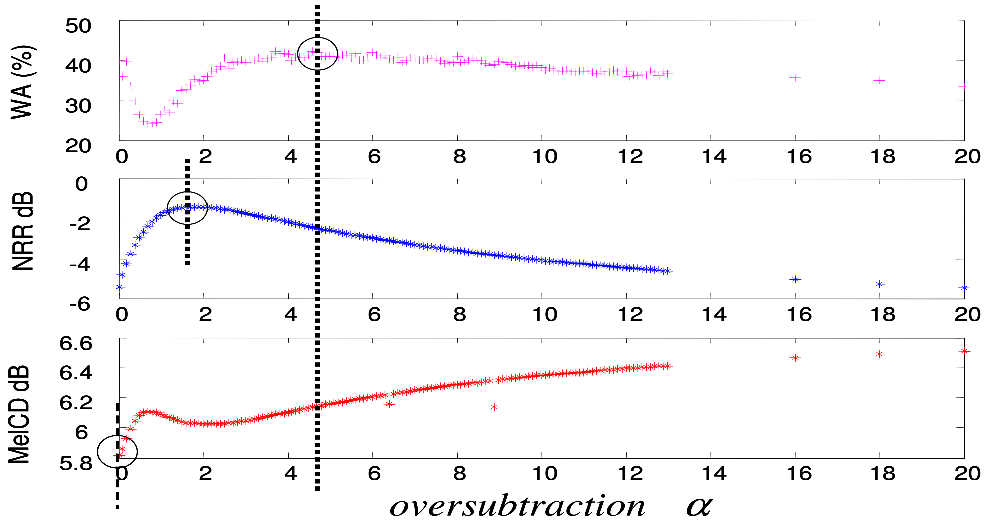


Figure 10. Plot of Word Accuracy, NRR, and MelCD in 10 dB office noise with varying α.

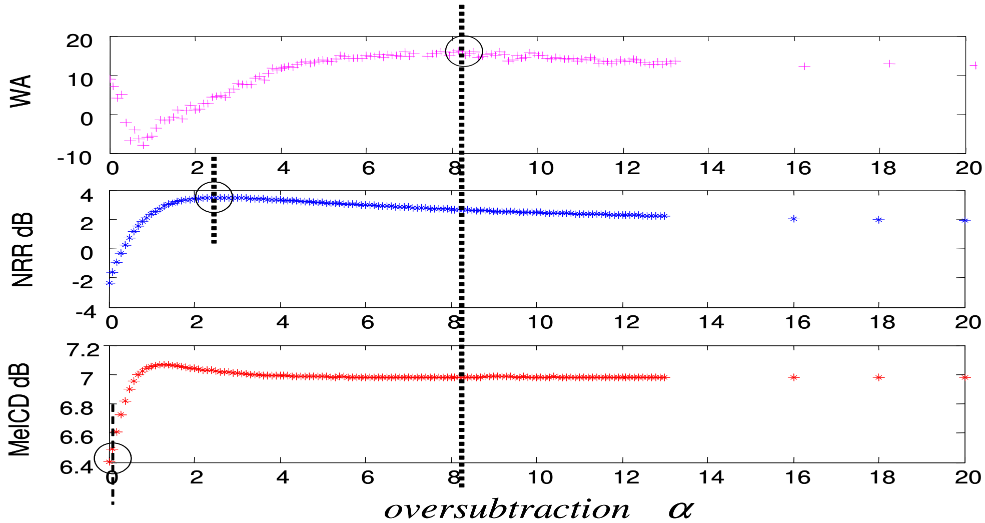


Figure 11. Plot of Word Accuracy, NRR, and MelCD in 0 dB office noise with varying α.

$$MelCD = \frac{20}{ln10}\sqrt{2\sum_{i=1}^{25}(mc_i^{orig}) - (mc_i^{new})^2}, \qquad (6)$$

where mc_i^{orig} and mc_i^{new} are the Mel Cepstrum coefficients before and after SS respectively. The NRR and MelCD give an indirect, yet informative insight of how the recognition performance might be. We investigate these parameters in parallel with the recognition performance as we denoise the test utterances with SS. Figures 9, 10, and 11 are the graphs showing the relationship of the recognition performance in Accuracy (word accuracy), NRR, and MelCD as a function of α in 25 dB, 10 dB, and 0 dB SNR respectively [12]. In Figure 9, where the original noisy speech utterances have SNR=25 dB, the correlation among Accuracy, NRR, and MelCD is well established. Maximum NRR and minimum MelCD correspond to the peaking of the word accuracy. Another result shows that, in a noisier condition with SNR≪25 dB as shown in Figures 10 and 11, Maximum NRR and minimum MelCD do not correspond to maximum accuracy in the same manner as that of SNR=25 dB. This result points to the fact how complicated it becomes

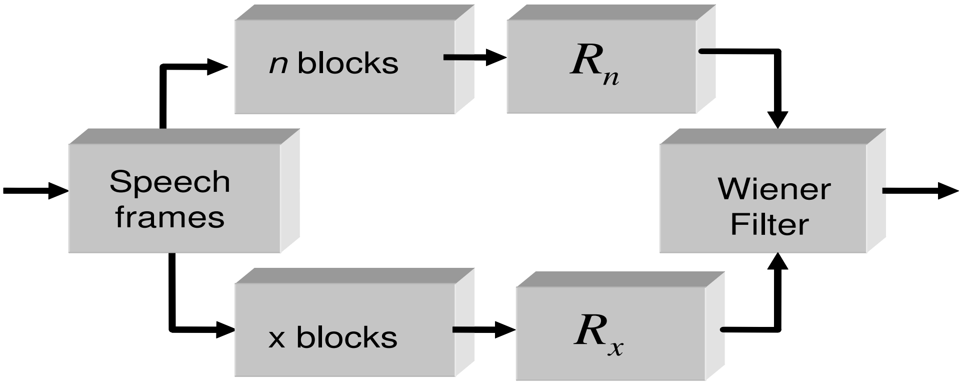


Figure 12. Block diagram of the time-domain Wiener filtering implementation.

when dealing with very low SNRs as far as the the recognizer is concerned. In a considerably high SNRs, speech enhancement like SS infers that a higher SNR and a lower distortion result to an improvement of the recognition performance. However as SNR decreases, this inference does not hold anymore in SS. Another speech enhancement technique is the *Wiener filtering* approach. Like many other denoising algorithms, this approach assumes that the interfering noise is additive and statistically independent which is simply written as

$$x = s + n, \tag{7}$$

where x,s, and n are the noisy signal, clean speech signal and the additive noise respectively. The figure of merit of Wiener filtering is that it gives the best estimate of the signal by minimizing the mean (expected value) of the squared error. There are many ways of implementing this approach. One example is the single channel time-domain implementation [13] shown in Figure 12. First, time domain noisy signal is divided into frames, after which the order of the filter is decided as given in Equation 8

$$L = round(\frac{20f_s + 2}{2}), \tag{8}$$

where L is the filter order and f_s is the sampling frequency in KHz.

The signal is then broken into blocks corresponding to the filter order, separating the noisy part (both the speech and the additive noise), and the noise-only part. The autocorrelation R_x is calculated for each of these blocks using Equation 9

$$R_x = \frac{\boldsymbol{x}^T \boldsymbol{x}}{L_S - L + 1}, \tag{9}$$

where L_s is the sample length. Wiener filtering is carried-out by first calculating the Wiener filter coefficients using the expression

$$\boldsymbol{W}_{WF} = R_x^{-1}(R_x - R_n), \tag{10}$$

where R_x and R_n are the autocorellation matrices of both the noisy signal and the noise-only signal. The denoised output Y is achieved by multiplying the Wiener filter coefficients to the transpose of the noisy blocked matrix given as

$$\boldsymbol{Y} = \boldsymbol{W}_{WF}\boldsymbol{x}^T. \tag{11}$$

Figure 13 shows the plot of the original signal corrupted by noise and the corresponding reconstructed signal using Wiener filtering. Multiple channels implementation of Wiener filtering can be found in the works of [14][15].

There are many different ways in reducing mismatch cause by noise. Denoising operations may be performed through combinations of various techniques, in different multi-stages and in different domains. The SS and Wiener filtering previously discussed are both operating in the time domain. The ETSI front-end approach implements hybrid enhancement techniques as shown in Figure 14. In this figure, denoising is done in different stages such as Wiener filtering and blind equalization in the ceptsrum domain. In Figure 15 a graph of the MFCC bins in a speech segment is shown when processed with ETSI together with the clean utterance itself without any denoising at all.

It should be clear by now, that denoising techniques implemented in one form or another have one purpose- to minimize the effect of mismatch. By removing the noise in the signal, we intend to preserve as much as possible the original state

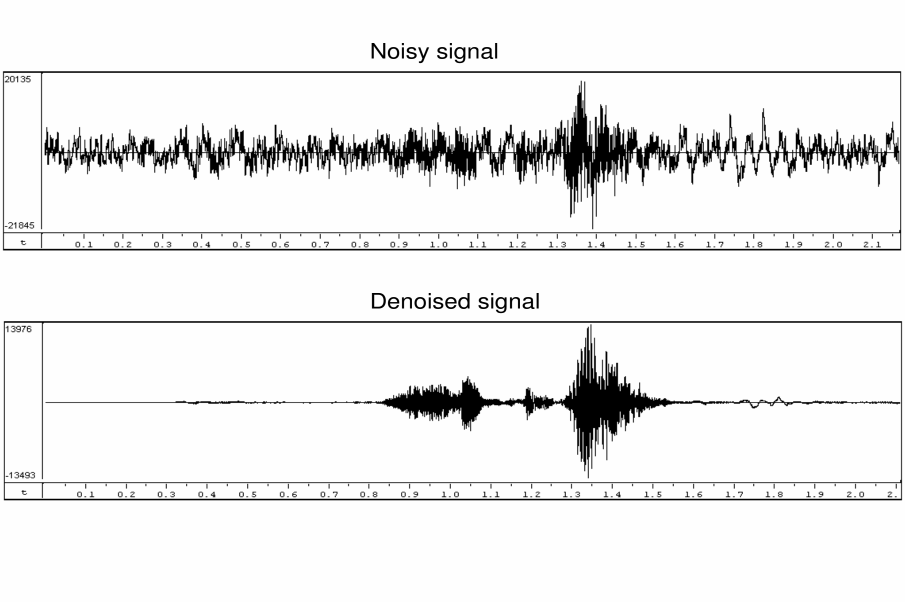


Figure 13. Original noisy signal and corresponding denoised signal using Wiener filtering.

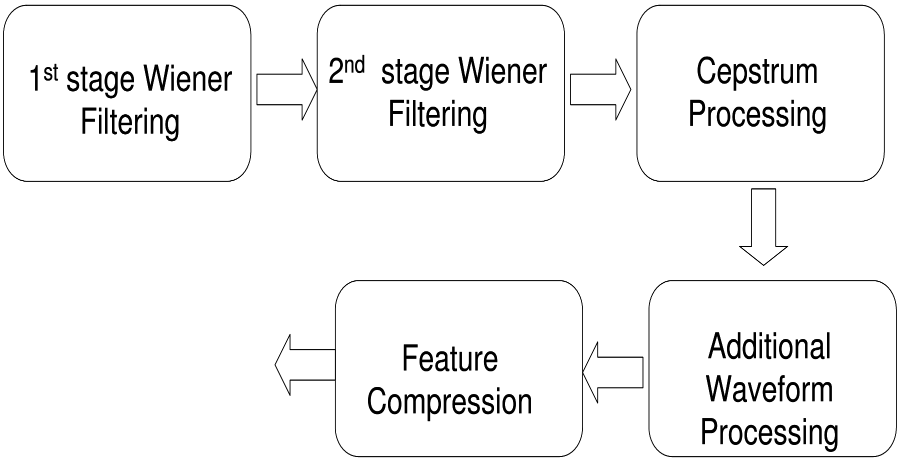

1st stage Wiener Filtering
2nd stage Wiener Filtering
Cepstrum Processing
Additional Waveform Processing
Feature Compression

Figure 14. Basic block diagram of the ETSI front-end noise robustness implementation.

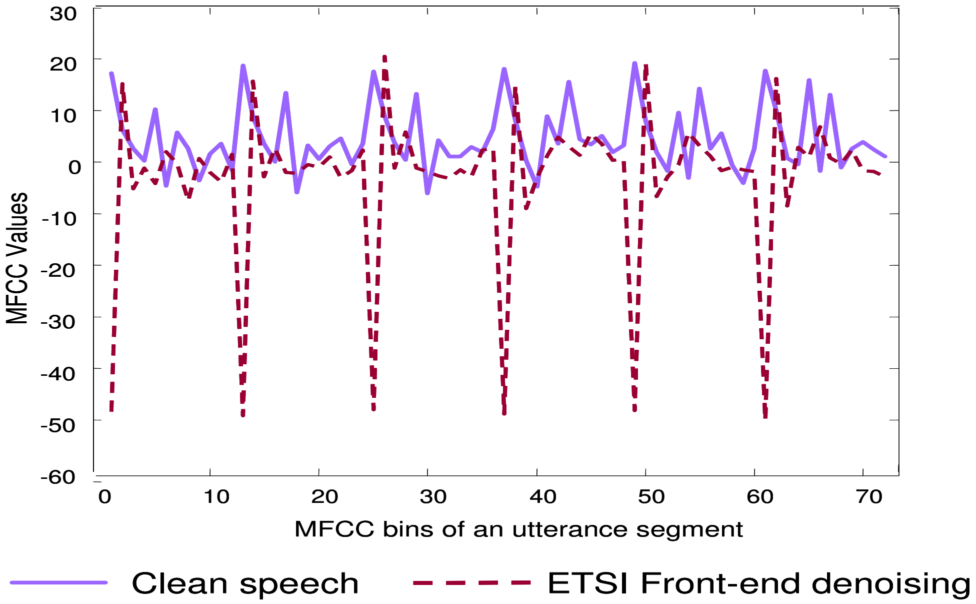


Figure 15. Plot of the MFCC of a speech segment when using ETSI and with no processing at all.

of the training utterances when creating the model. This is the very same model we use in performing recognition given the noisy test utterance. This however, can be solved by creating different models matched with different acoustical environments, but there are so many scenarios of different acoustic environments that we can think of, thus creating matched models is impractical if not impossible. Later in this thesis we will show our approach to this problem. In getting away of matching models with different types of noise, we introduce a robust model.

It has been shown how complicated the problem becomes under noisy environment conditions. The basic SS recognition experiment for example, shows that no matter how much we attempt to correlate the recognition performance with different measures such us NRR and MelCD, these become more unreliable as SNR decreases.

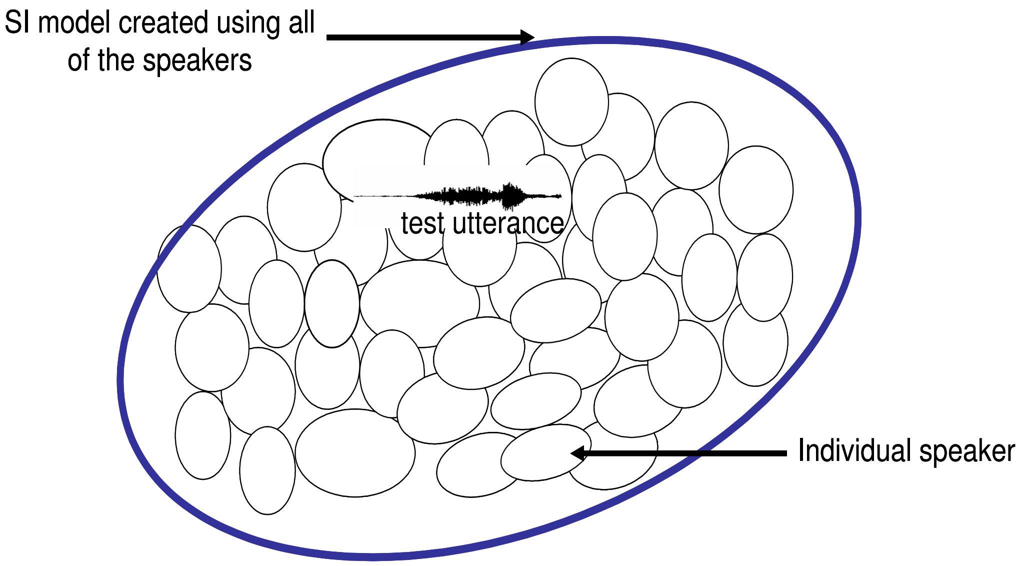


Figure 16. Visualization of a single-template model.

2.5.2 Speaker Normalization and Adaptation

An accurately trained SI model is possible by means of using multiple training database of various genders and age-groups as shown in Figure 16. Although it reduces the sensitivity to different genders due to a broader age and gender spectrum, this approach is likely to render the SI model to have an increase in variance due to the wide varieties of speakers. Net effect is, recognition distribution would be flat due to the averaging of too many speakers [16]. Studies in speaker variability show that there are several methods in addressing this problem [17]. One way of dealing with speaker variability is to use some training techniques like the cluster-based modeling approach which results to an improvement in recognition performance by training multiple classes of acoustic models with smaller variances together with an appropriate model selection method [18].

Pre-processing techniques during feature extraction prior to modelling are also widely used like the Vocal Tract Length Normalization VTLN [19] [20] which effectively compensates the different sizes of speakers' vocal tracts through frequency warping. Experiments in adult and children data yield an improvement in recognition accuracy when using VTLN [21].

Another method in minimizing the effect of speaker variability is to employ model adaptation in which our proposed method falls in the same category. This approach effectively adjusts the SI model to reflect the inherent characteristics of the adaptation data to the adapted model. Popular techniques that belong to this category are the Maximum Likelihood Linear Regression (MLLR) [22] and Maximum A Priori (MAP) [23]. Model adaptation by means of transformation and combination of HMMs [24] and smooth N-best based speaker adaptation [25] are also proposed.

2.5.3 Minimum data and Rapid Approach

To address both adaptation data and adaptation time issues in speaker adaptation, it is important to find ways in which adaptation algorithms are responsive even with very small adaptation data. Moreover, computational load should be kept minimal. Works like the linear combination of rank-one matrices [26] and a very fast compact context-dependent eigenvoice model adaptation [27] that can handle short adaptation data and adapt fast are presented. Unsupervised speaker adaptation based on HMM-Sufficient Statistics has been proposed [28]. This is a promising approach for a fast adaptation using only one arbitrary adaptation utterance without transcription. In this thesis we will discuss the concept and the development of this rapid adaptation scheme.

2.6 A Closer Look On Mismatch

Mismatch is not just caused by additive noise nor the immense variability of the speech signal itself. Different systems have different configurations and this can cause mismatch. Consistency in processing the speech is very important as these are reflected in the model. An example of mismatch in the parameterization process are shown in Figures 17 and 18. These figures show the spectra of the vowel “i” and “o” processed independently with SS and the ETSI front-end. These discrepancies affect the model being trained that will render one model to be completely not usable with the other processing or the other way around. Another illustration showing mismatch in the MFCC for a single utterance is

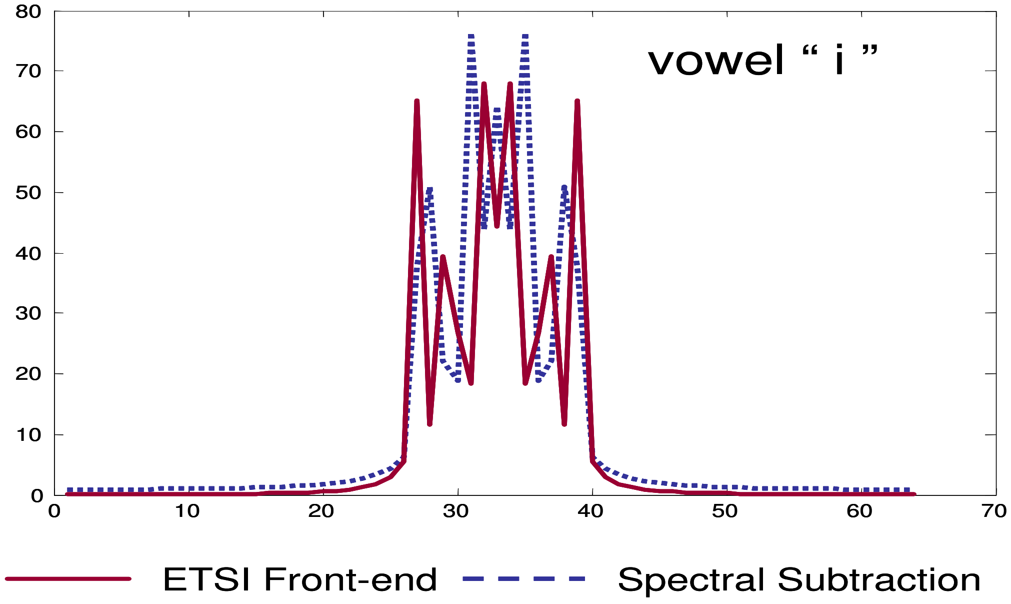


Figure 17. Discrepancies in the spectrum of the Japanese vowel “i” processed by SS and ETSI front-end.

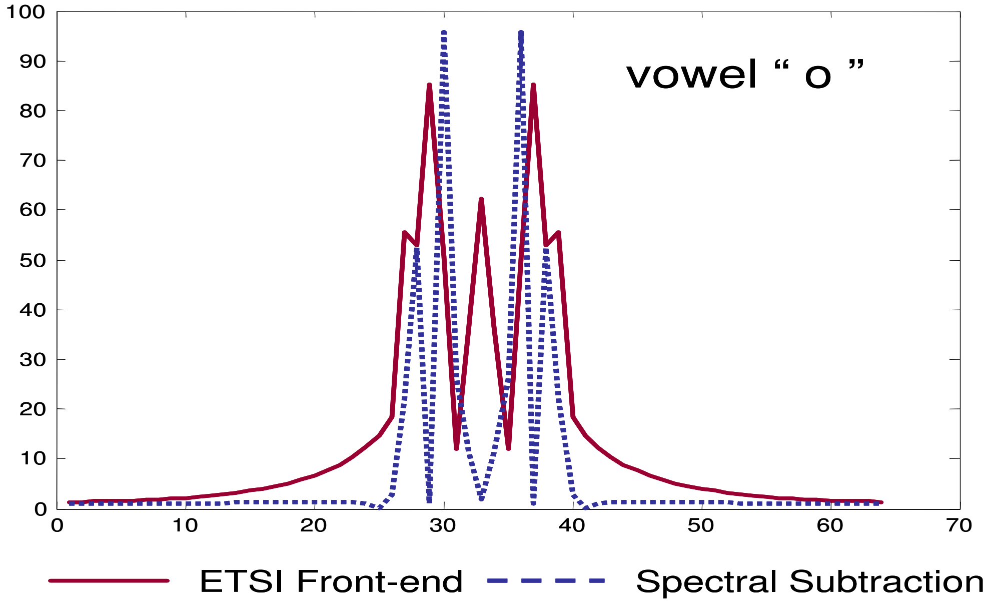


Figure 18. Discrepancies in the spectrum of the Japanese vowel “o” processed by SS and ETSI front-end.

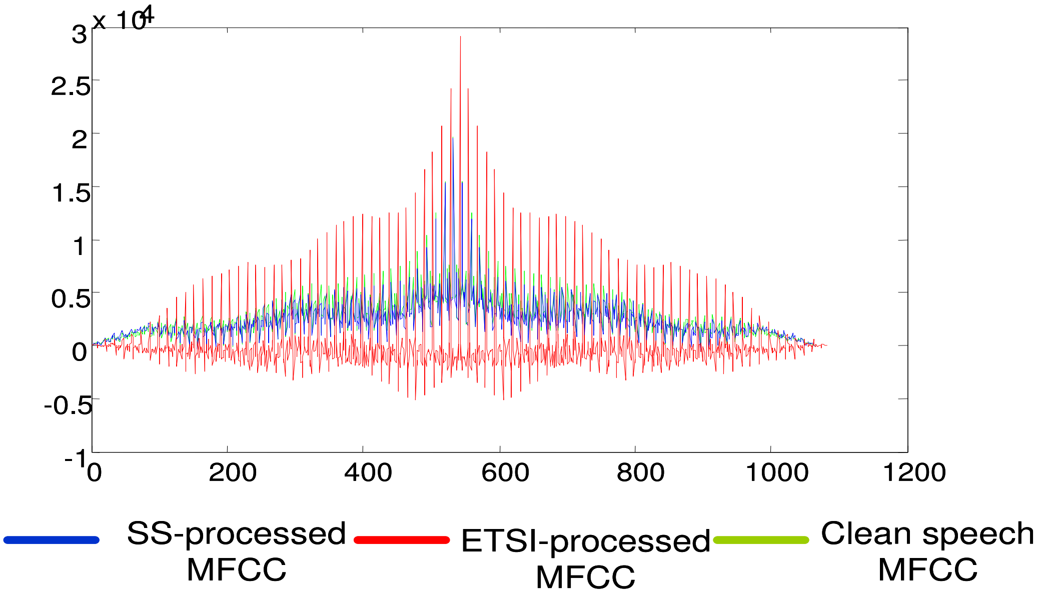


Figure 19. Mismatch in the MFCC values of a single utterance when using different processing.

shown in Figure 19.

2.7 The Proposed Approach

We have discussed the problems that degrades the performance of the ASR in general context, and the available approach needed to minimize its effects. We will introduce the merits of our approach towards a robust speech recognition in real environment conditions. In later chapters, a more detailed explanation will be presented.

2.7.1 Under Real Environment Condition

Our proposed approach is designed to work in noisy environment conditions. By superimposing 25 dB office noise to the utterances prior to training and superimposing 30 dB office noise to the noisy utterance denoised by SS, a single robust model is created to avoid the use of matched models with different types of noise. Moreover, we evaluate the systems performance in car, crowd, booth, and office

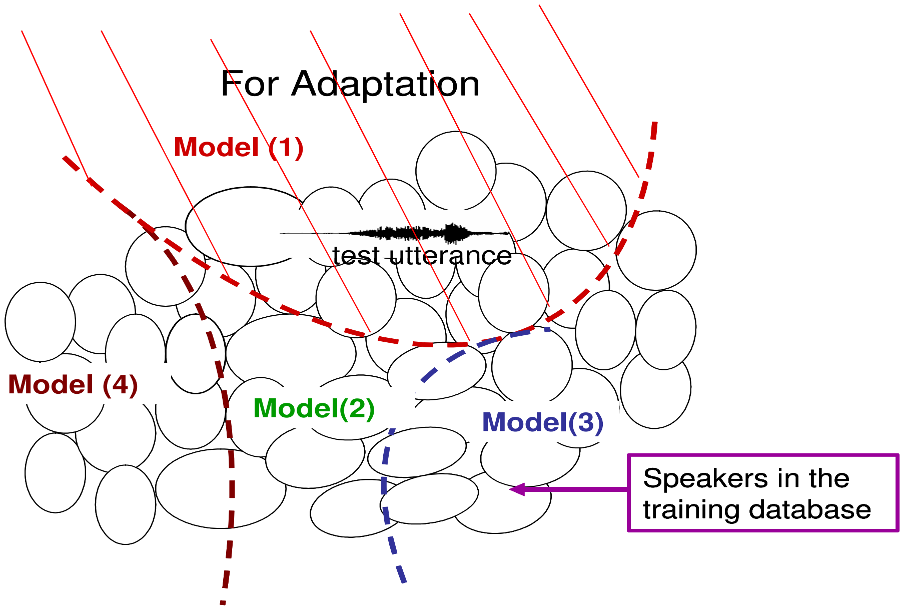


Figure 20. Using multi-template selection and model adaptation for robustness in speaker variation.

environment with different SNRs.

2.7.2 Robustness in Speaker Variation

The proposed adaptation method deals the problem of speaker variation with two series of improvements. First, we employ multi-template model selection based on the acoustic similarity of the test utterance to improve the performance of the system as opposed to using only a SI model. Second, to further improve the acoustic model relative to the test utterance, adaptation is performed using the selected model and the HMM-Sufficient Statistics as adaptation data. This is illustrated in Figure 20 where “S.I.(1)” is selected based on its acoustic similarity with the test utterance and then adaptation is performed using the selected model and the corresponding adaptation data which belongs to the class of the selected model.

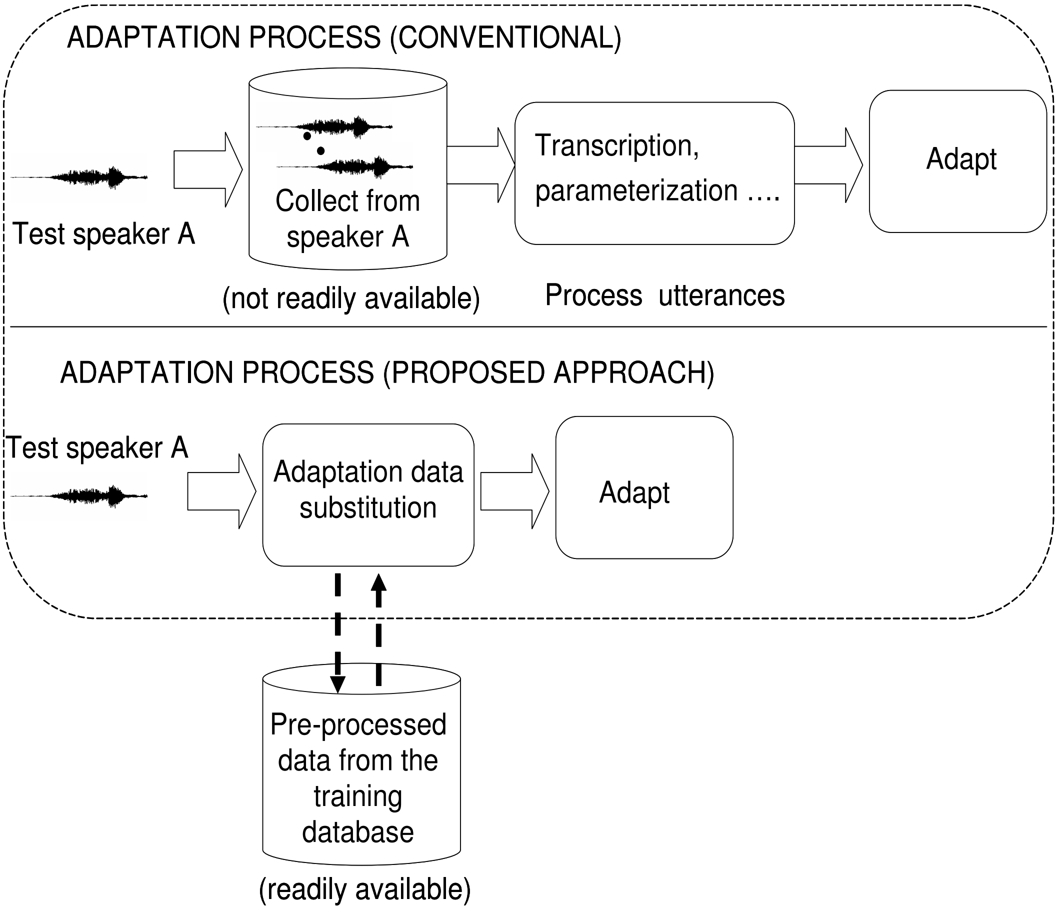

ADAPTATION PROCESS (CONVENTIONAL)
Test speaker A
Collect from speaker A
(not readily available)
Transcription, parameterization ….
Process utterances
Adapt
ADAPTATION PROCESS (PROPOSED APPROACH)
Test speaker A
Adaptation data substitution
Adapt
Pre-processed data from the training database
(readily available)

Figure 21. Overview of the proposed recognition approach using only a single arbitrary utterance for adaptation.

2.7.3 Rapid Adaptation Using Single Adaptation Utterance

Adaptation data in the form of Sufficient Statistics are pre-parameterized and stored during the training process. By making use of discriminate selection of speakers, we are able to employ a mechanism replacing the actual adaptation data with the HMM-Sufficient Statistics making adaptation faster and more efficient. Figure 21 compares the conventional adaptation and the proposed rapid adaptation using a single utterance only. In the conventional approach, in order to adapt to the test utterance A, data from speaker A is collected in many utterances to serve as an adaptation data. This process cannot be done before hand. The collection takes place on-site which takes a lot of time. When adaptation data collection is finished, supervised transcription is needed together with the parameterization of the speech utterances into suitable format. Until then, the actual adaptation commences. In the bottom part of the figure, the proposed method is shown. It is apparent that data collection is not necessary since the method has a mechanism of substituting the needed adaptation data with some replacements, this process will be explained later in details. For now, we assume that adaptation data is readily available to the system anytime for adaptation. The proposed approach, only uses a single arbitrary utterance for adaptation and since adaptation data is readily available, time-consuming processes are by-passed which is absent in the conventional approach.

2.8 Summary

In this chapter, we focused on the external factors that affect the performance of the recognizer. We have discussed in detail the two most important problems that arise in real environment conditions namely, additive noise and speaker variability. Examples in forms of illustrations are given to understand the nature of these problems and how these two impact the performance of our recognizer. It is noted earlier that recognizers tend to perform poorly as the SNR increases or as it becomes noisier. Aside from the additive noise, speaker variability also degrades performance of the system. We reviewed current research that deals with these problems and implementation towards a more robust speech recognition system. We then introduced briefly how our approach would solve or minimize the effects of these problems in a very practical manner. In Chapter 3, we will discuss the basic concepts in statistics that will lead to the theoretical framework of the HMM-Sufficient Statistics in Chapter 4. Chapters 5 and 6 will discuss the two adaptation schemes based on HMM-Sufficient Statistics, the rapid Baum-Welch reestimation and the rapid MLLR adaptation. Results will be discussed in detail in Chapter 8 while in Chapter 9 we will evaluate the performance of the proposed method in real environment conditions, being integrated in an actual dialogue system. We conclude this thesis in Chapter 9.

3. Statistics Theory in Speech Recognition and the Proposed Rapid Unsupervised Speaker Adaptation

In designing classifiers, knowledge of the probabilistic structure of the problem is very important but unavailable in a real pattern classification applications. We resort to train a classifier using the limited information within our disposal. Two common approaches are available namely, Bayesian and Maximum Likelihood estimation. In this chapter, the basic statistical tools are introduced which are used in gradually establishing the theoretical concept of the proposed adaptation technique in Chapter 4.

3.1 Bayes' Decision Theory

A basic approach to the parameter estimation problem is the Bayesian decision theory. In this theory, decision is achieved by acknowledging that the given classification problem is probabilistic in nature and comes along with it is the cost in every decision to make. First, we need to identify the states that we are interested for a particular classification problem. The states simply represent the classes needed to be identified. We denote the state as ω_n where n is the *nth* class. Moreover, in the Bayesian approach, it is important to know in advance some priors which actually affects the outcome of the next state. It is referred to as the *a priori probability* denoted by $P(\omega_n)$ s.t. for N number of classes we impose

$$\sum_{n=1}^{N} P(\omega_n) = 1. \tag{12}$$

In classification problems, there are several variables available that help gather more information pertaining to a certain class. If tasked to solving a classification problem, it is important to identify these variables and refer to these from time to time to establish a measurement of similarity or dissimilarity between objects of interest. This becomes the basis for classification. Suppose that variable x gives a measurement of a certain property present in all objects to be classified, but varies accordingly in each of the objects. This information can be used and expressed in a probabilistic manner to improve the performance of the classifier. The *class conditional probability density function* $p(x|\omega_n)$ is introduced, this is referred as the *likelihood* of ω_n. Suppose x is some measurement of a certain property, then $p(x|\omega_1)$ and $p(x|\omega_2)$ show the individual measurements in classes ω_1 and ω_2 respectively. Since, prior information of this measurement is assumed to be available in every object needed to be classified, the *class conditional probability density function* gives an additional information of how much measurement of a certain property is present in the unknown object.

After putting all these probabilistic concepts, the state of an unknown object can be meaningfully predicted. The *joint probability density* $p(\omega_n, x)$ is given as

$$\begin{aligned} p(\omega_n, x) &= P(\omega_n|x)p(x) \\ &= p(x|\omega_n)P(\omega_n). \end{aligned} \tag{13}$$

By using 13, the question of how the measurements x affect the identification of the actual class of the unknown object can be answered. From Equation 8, the Bayes' formula is expressed as

$$P(\omega_n|x) = \frac{p(x|\omega_n)P(\omega_n)}{p(x)}, \tag{14}$$

where $p(x|\omega_n)$ and $P(\omega_n)$ are the *conditional density function* and the *prior probability* respectively. The denominator $p(x)$ functions as a normalizing factor given by

$$p(x) = \sum_{n=1}^{N} p(x|\omega_n)P(\omega_n). \tag{15}$$

The Bayes' formula in Equation 14 shows that through the observation of the value x as given by the *likelihood*, the *prior* can be converted to the *posterior* which is the probability of the state of nature of being in the class ω_n given the measurement x. From these onwards, it is obvious that decision can be made using the result of the posterior. For two classes ω_1 and ω_2, and with an observation x, the class that results to a higher posterior value is more likely the actual class to be classified. The *probability of error* is introduced in order to establish a decision rule. A simple expression for the probability of error whenever x is observed is

$$P(error|x) = min[P(\omega_1|x), P(\omega_2|x)]. \tag{16}$$

The posterior that results to a minimum *probability of error* influences the decision of identifying the unknown class.

In real classification problem, it is often expected to use more than one feature. Thus Equation 14 is expressed as

$$P(\omega_n|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_n)P(\omega_n)}{p(\boldsymbol{x})}, \tag{17}$$

and $p(\boldsymbol{x})$ is given as

$$p(\boldsymbol{x}) = \sum_{n=1}^{N} p(\boldsymbol{x}|\omega_n)P(\omega_n). \tag{18}$$

Moreover, a loss function can be introduced in decision rule instead of using the *probability of error*. By defining that a loss $\psi(\delta_i|\omega_n)$ is incurred whenever taking an action δ_i if the true state of nature is ω_n. The accumulated loss is

$$R(\delta_i|\boldsymbol{x}) = \sum_{n=1}^{N} \psi(\delta_i|\omega_n)P(\omega_n|\boldsymbol{x}). \tag{19}$$

where $R(\delta_i|\boldsymbol{x})$ is called as the *conditional risk*. It is apparent that in order to minimize the expected loss whenever $\boldsymbol{x}$ is encountered, the corresponding action selected should minimize Equation 19 the *conditional risk*. The introduction of the *conditional risk* paves the way of calculating the *overall risk*. The problem now evolves in finding a decision rule that minimizes the *overall risk*.

Our objective is to look for δ_i that would minimize $R(\delta_i|\boldsymbol{x})$. This will serve as a decision rule in which action to take given every possible observation against

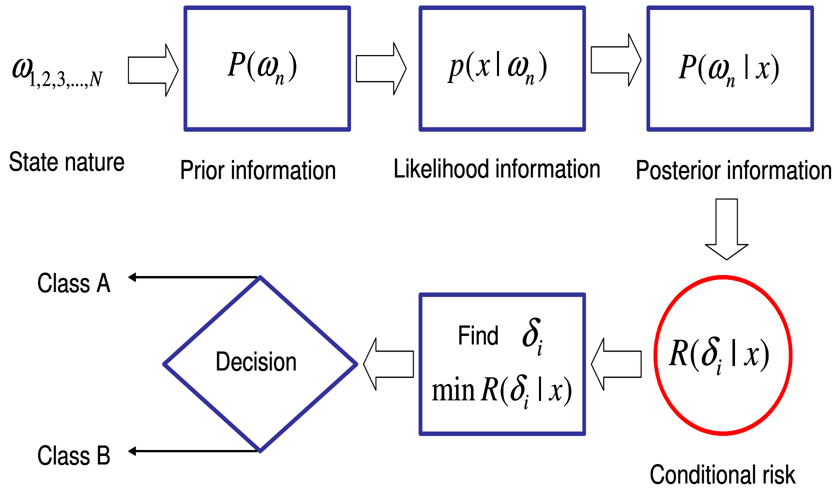


Figure 22. Block diagram of basic Bayesian decision rule.

$P(\omega_n)$ such that the overall risk is minimized. Figure 22 summarizes the basic steps in identifying classes using the basic decision rule. State nature $\omega_{1,2,3,\ldots,n}$ is identified beforehand, which is basically identifying the possible classes involved in the problem. It is assumed that the *prior probability* is known, and by using the likelihood information, the *prior probability* is converted into *a posterior probability*. Consequently, the *conditional risk* is associated with every action δ_i. A corresponding δ is selected which minimize the *conditional risk*. Finally, decision rule is established based on the minimum risk.

An illustration of classifying two sets of classes and the advantage of using Bayes' approach is shown in Figures 23 and 24. Figure 23 shows a relatively simple classification problem where the decision boundary is easily solved linearly. This kind of classification problem occurs most often when there are two classes involved and when these classes are very dissimilar to each other. In Figure 24 however, we see a more complicated picture where it is difficult to separate the boundaries between two classes linearly. We tried a simple Least-Square algorithm to establish the boundaries between class A and class B and obviously, LS results a poor separation between class A and B. However, if we resort to using Bayes' decision rule, the separation between classes A and B as shown by the dashed line is better compared to LS. This shows the advantage of using Bayes'

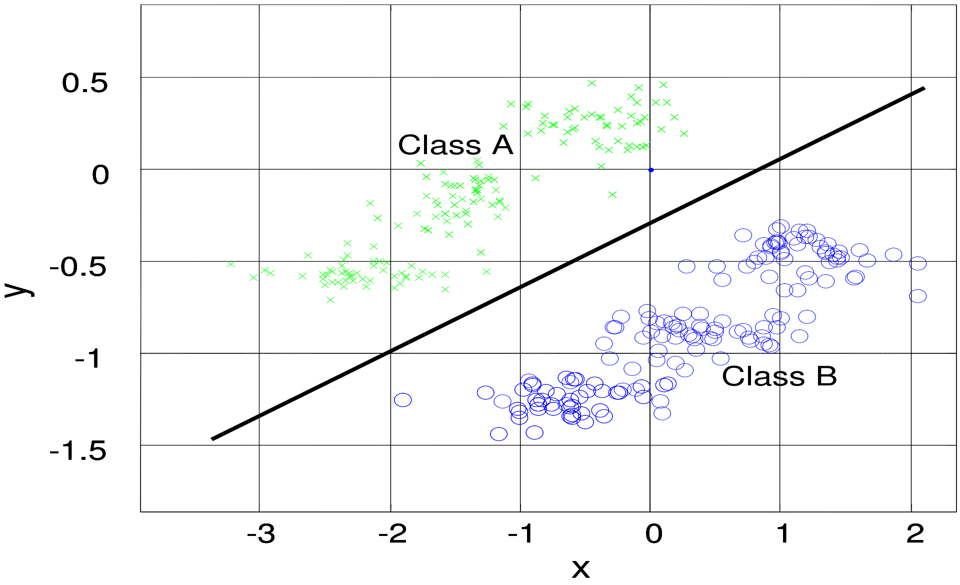


Figure 23. An example of an easy pattern classification problem where two classes can be linearly separated.

decision rule in classifying patterns that cannot be easily linearly separated.

3.2 Bayesian Estimation

Bayesian learning or Bayesian Estimation approach does not assume $\boldsymbol{\theta}$ to be fixed given a parametric form $Z(\boldsymbol{\theta})$. It treats this as a random variable and together with the training data, finds a solution of an unknown distribution. In Figure 25, the process of Bayesian estimation is explained. The objective is to find the unknown probability density $p(\boldsymbol{x})$ but since this is not feasible, we resort to solving for $p(\boldsymbol{x}|D)$. This is done by calculating for $p(\boldsymbol{x}, \boldsymbol{\theta}|D)$. Thus solving for $p(\boldsymbol{x}|D)$ is actually dependent on $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|D)$. Note that these two are both known. To be more specific, we need to integrate the joint density $p(\boldsymbol{x}, \boldsymbol{\theta}|D)$ over θ and get

$$p(\boldsymbol{x}|D) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|D) d\boldsymbol{\theta}. \tag{20}$$

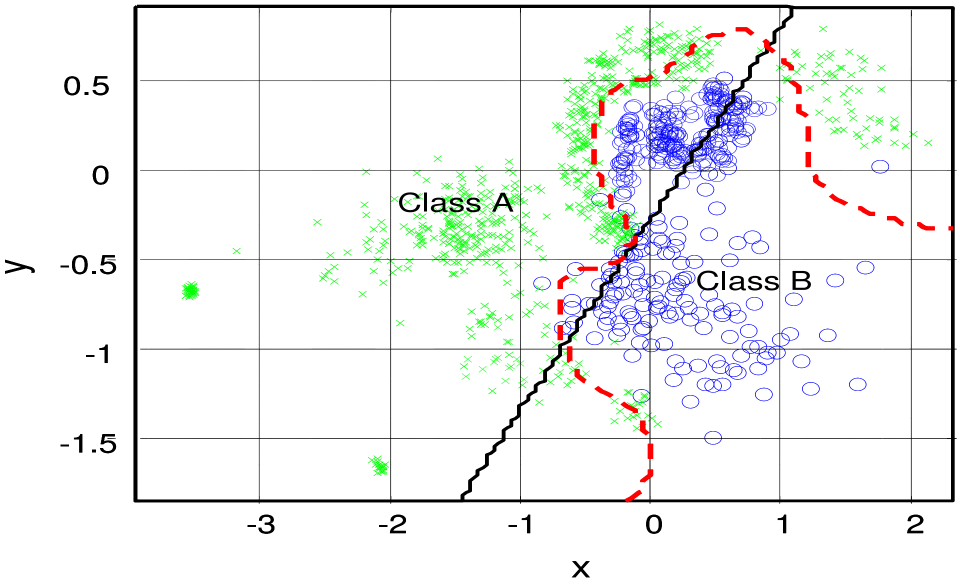


Figure 24. Comparison between Bayes' and LS when classifying two classes whose boundary cannot be separated linearly.

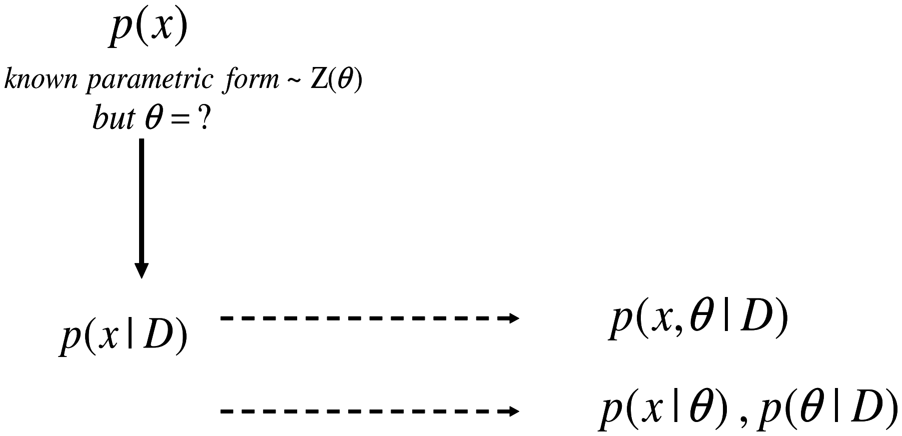

$p(x)$
known parametric form ~ $Z(\theta)$
but $\theta = ?$
$p(x \mid D)$
$p(x, \theta \mid D)$
$p(x \mid \theta), p(\theta \mid D)$

Figure 25. The Bayesian estimation method.

But

$$p(\boldsymbol{x}, \boldsymbol{\theta}|D) = p(\boldsymbol{x}|\boldsymbol{\theta}, D)p(\boldsymbol{\theta}|D), \tag{21}$$

since both $\boldsymbol{x}$ and D are independently selected, Equation 20 becomes

$$p(\boldsymbol{x}|D) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}. \tag{22}$$

3.3 Maximum Likelihood Estimation

Optimal classifiers can be designed if information such as the *prior probabilities* $P(\omega_i)$ and *class conditional probabilities* $p(\boldsymbol{x}|\omega_i)$ are available. However, we do not have access to these information, and the least information available to us is the general knowledge of the situation and representations of the patterns that we need to classify. If the parameters are identified beforehand, and if the *class conditional probabilities* $p(\boldsymbol{x}|\omega_i)$ can be parameterized by using the general knowledge about the data, then Maximum Likelihood Estimation (MLE) can be used. It is reasonable to assume that $p(\boldsymbol{x}|\omega_i) \backsim Z(\boldsymbol{\theta})$, Hence, the problem of estimating $p(\boldsymbol{x}|\omega_i)$ settles down to estimating the parameter $\boldsymbol{\theta}$ for the known density $Z(\boldsymbol{\theta})$. In the case of the Normal distribution, $\boldsymbol{\theta}$ is consist of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In this context we refer to *likelihood* of the parameter to be estimated with respect to the training data

$$p(D|\boldsymbol{\theta}) = \prod\nolimits_{k=1}^{n} p(\boldsymbol{x}_k|\boldsymbol{\theta}), \tag{23}$$

where D is the set of training data containing n samples $\boldsymbol{x}_1$,...,$\boldsymbol{x}_n$ and $\boldsymbol{\theta}$ is the vector-valued parameter to be estimated. We define

$$\phi(\boldsymbol{\theta}) \equiv p(D|\boldsymbol{\theta}). \tag{24}$$

The ML estimate $\widehat{\theta}$ is the value of θ that maximizes the likelihood $p(D|\boldsymbol{\theta})$ ie.

$$\widehat{\boldsymbol{\theta}} = arg \underset{\boldsymbol{\theta}}{max} \phi(\boldsymbol{\theta}) \tag{25}$$

For analytical purposes, it is preferable to work in terms of log-likelihood rather than the likelihood itself. Since the logarithm is monotonically increasing, the parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood also increases the likelihood itself. Thus, the expression

$$\log \phi(\boldsymbol{\theta}) = \sum\nolimits_{k=1}^{n} \log p(D|\boldsymbol{\theta}), \tag{26}$$

and to maximize θ we make sure that

$$\frac{\partial \log \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0. \tag{27}$$

As noted earlier, MLE is important in establishing the significance of the proposed adaptation method based on HMM-Sufficient Statistics. Hence, a very brief example using ML estimation is discussed. Suppose that samples are drawn from a multivariate normal population having a mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. By considering sample point $\boldsymbol{x}_k$ we get

$$\ln p(\boldsymbol{x}_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] - \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\theta}). \tag{28}$$

For simplicity, we only consider the case where $\boldsymbol{\Sigma}$ is given, thus we only need to find for $\boldsymbol{\mu}$. Thus

$$\ln p(\boldsymbol{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] - \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}). \tag{29}$$

To find $\widehat{\boldsymbol{\mu}}$ we use Equation 27. Since $\boldsymbol{\Sigma}$ is assumed to be given, the only variable in $\boldsymbol{\theta}$ is $\boldsymbol{\mu}$. Thus, we maximize the log-likelihood with respect to $\boldsymbol{\mu}$ such that

$$\frac{\partial \ln \phi(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 0. \tag{30}$$

The ML estimate of $\boldsymbol{\mu}$ must satisfy

$$\frac{\partial}{\partial \boldsymbol{\mu}}(-\frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] - \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu})) = 0. \tag{31}$$

and obtain $\widehat{\boldsymbol{\mu}}$ as

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k\ . \tag{32}$$

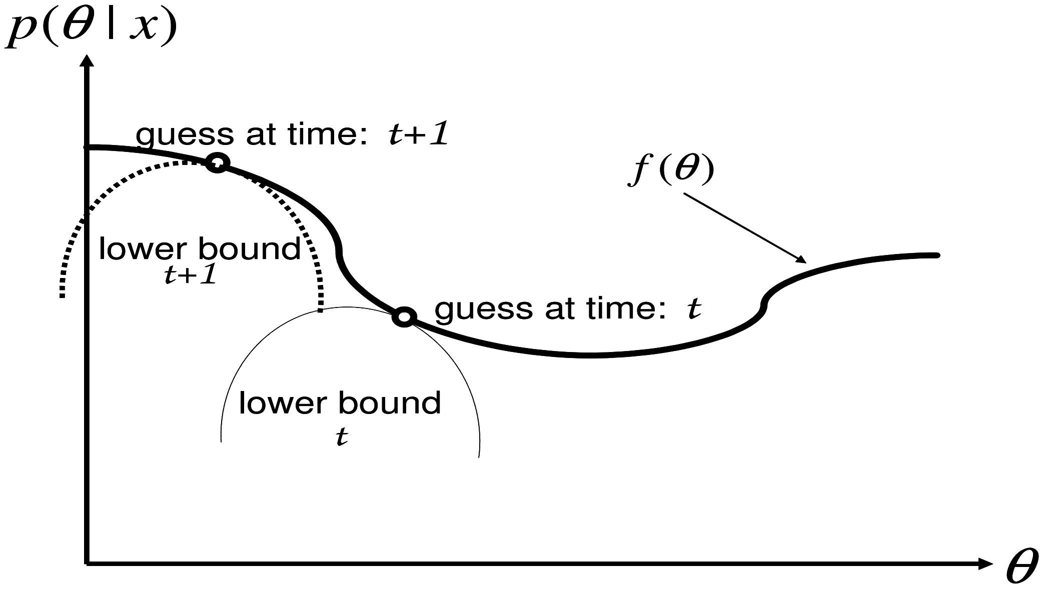


Figure 26. Overview of the E-M algorithm maximizing the objective function with a lower bound.

This result implies that the sample mean which is the average of the training data is a ML solution. We will discuss later the relevance of this result in connection with Sufficient Statistics.

3.3.1 Expectation Maximization

As a solution to the ML problem, an iterative technique is used in estimating probabilistic model parameters $\boldsymbol{\theta}$ given some data D [29] [30] [31] [35] [36]. This optimization scheme is also called as *primal dual* method [32] [33] [34]. A graphical illustration of the optimization process used in E-M is shown in Figure 26, where an iterative lower bound estimate $B(\theta_t, \theta_{t+1})$ touches the objective function $f(\theta)$ [37] [38] [39] [40]. Moreover, as it proceeds from current guess at time t to $t = t + 1$ the current lower bound estimate is always a better estimate of the previous one unless there is no change in the gradient at that particular point. The overall cycle of lower bound estimates would lead to the maximization of the objective function $f(\theta)$. Thus clearly, the whole process is composed mainly of

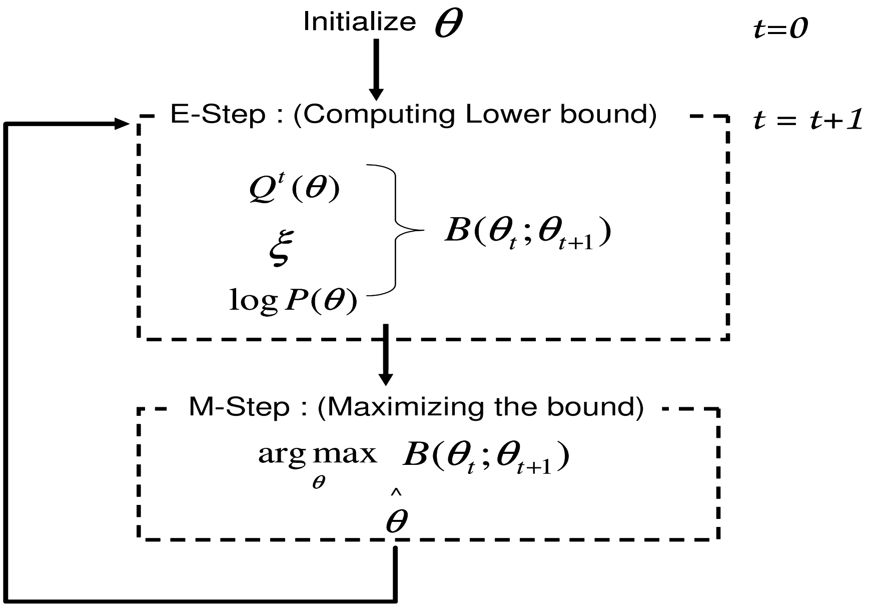

Initialize θ
t=0
E-Step : (Computing Lower bound)
t = t+1
$Q^t(\theta)$
ξ
$\log P(\theta)$
$B(\theta_t;\theta_{t+1})$
M-Step : (Maximizing the bound)
$\arg\max_{\theta} B(\theta_t;\theta_{t+1})$
$\hat{\theta}$

Figure 27. Basic illustration of the iterative E-M Algorithm used to find and maximize the bound.

calculating the lower bound which is called the “E-step” and the “M-step” which is maximizing the lower bound in order to touch the objective function.

The algorithm is summarized in Figure 27 where an initial estimate $\boldsymbol{\theta}$ is used at $t = 0$. For the E-step which is comprised of calculating the lower bound $B(\theta_t, \theta_{t+1})$, we calculate the *expected log-likelihood* $Q^t(\boldsymbol{\theta})$, the *entropy* $\boldsymbol{\xi}$, and the log of $P(\theta)$. To make sure that the lower bound touches the objective function, the bound is maximized through the M-step. The resulting $\widehat{\boldsymbol{\theta}}$ is used as input to the E-step and the process is iterated. The E-M algorithm sometimes result to a faster convergence as compared to the Newton’s method [41]. However, this is still a local technique which suffers local minima problems. Thus, it is very important to have a better initial estimate at the beginning to decrease the chances of being stuck at a local minima. The concept of a good estimate is important to understand the advantage of multi-template models we used in our proposed adaptation in the later chapters. An illustration of the impact of using a good initial estimate for the E-M are given in Figures 28- 31 which are the $1st$, $2nd$, $3rd$, and $4th$ (last) iterations respectively. We used a good estimate for θ

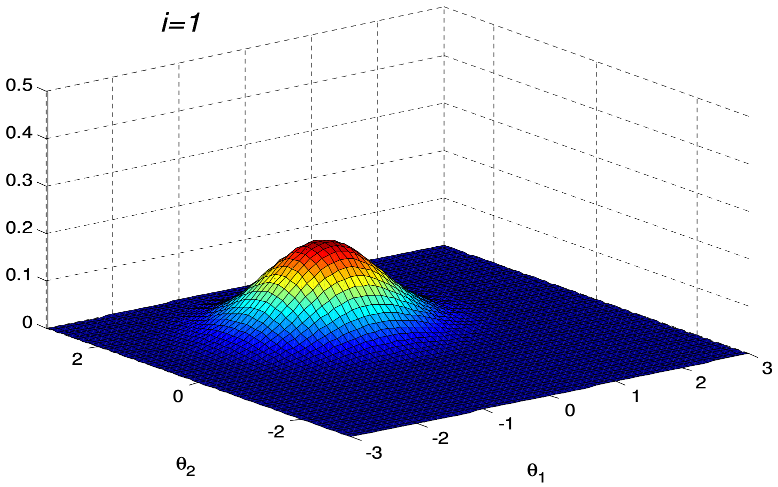


Figure 28. The lower bound estimate that touches the likelihood function with iteration 1 using a good initial estimate.

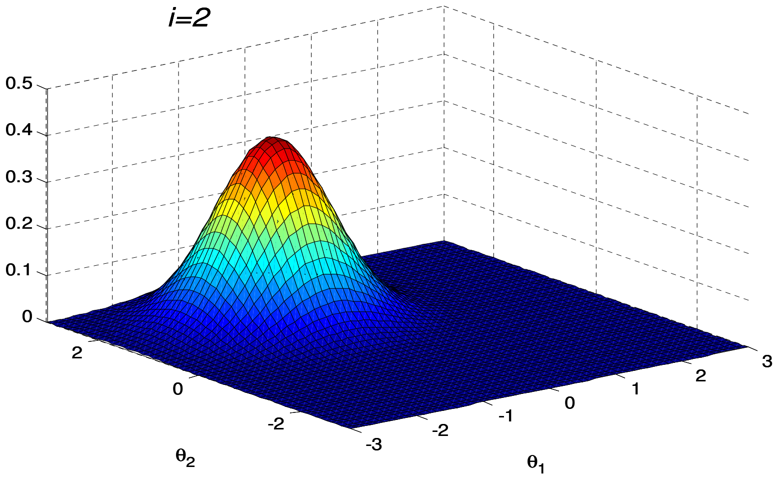


Figure 29. The lower bound estimate that touches the likelihood function with iteration 2 using a good initial estimate.

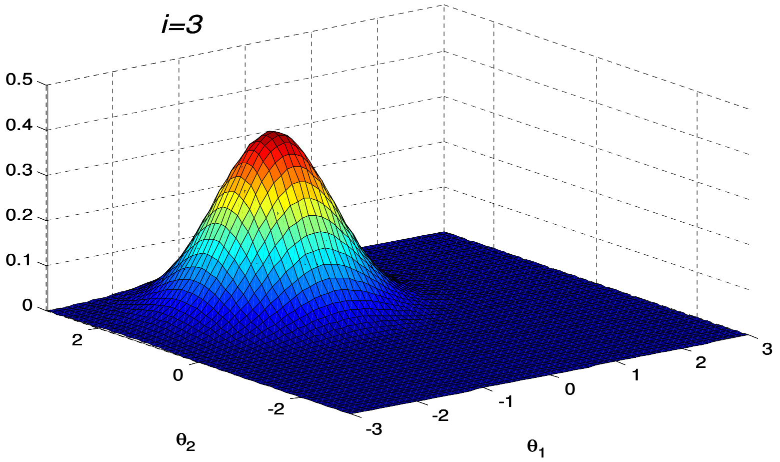


Figure 30. The lower bound estimate that touches the likelihood function with iteration 3 using a good initial estimate.

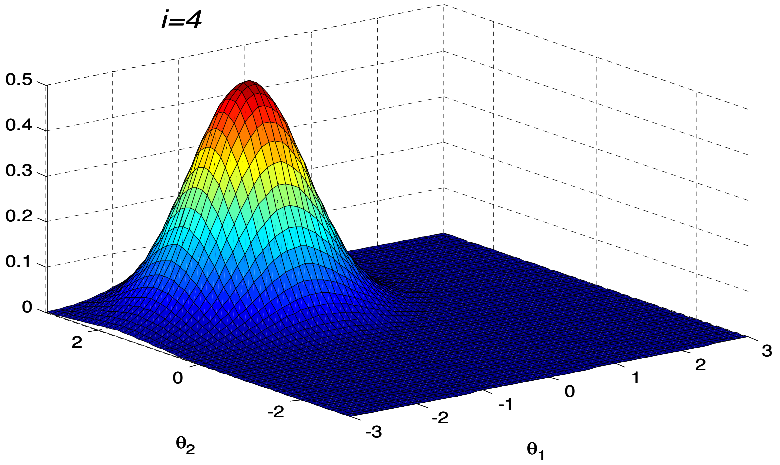


Figure 31. The lower bound estimate that touches the likelihood function with iteration 4 using a good initial estimate.

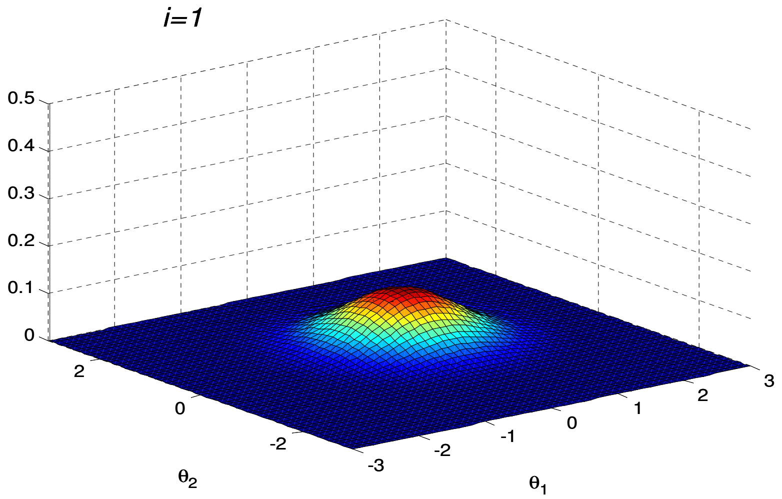


Figure 32. The lower bound estimate that touches the likelihood function with iteration 1 using a random initial estimate.

at the very beginning thus it takes only four iteration in order to complete the whole process in finding $\widehat{\theta}$. On the other hand, in Figures 32 - 37 when using E-M with the same training data but using random initial estimates. In these figures, it is apparent that it takes two more iterations to complete the whole process as opposed to using a good initial estimate in Figures 28 - 31.

Another important consideration when dealing with E-M for mixture model is the size of the training database and the number of mixtures. As a rule of thumb, it is better to increase the number of mixtures when the training database is of considerable size. That is, the more training data we have, the performance would significantly improve if we increase our mixtures as a result of a better model fit. However, increasing the mixtures indiscriminately would result to a degradation of the performance due to insufficiency of training data. This is an important introductory concept since in later chapters we will be dealing with models with much more than single mixture (i.e. 64 Gaussian mixtures). As an illustration, E-M is used to separate two classes of data. Figure 38 shows the performance when using only a single Gaussian mixture. We increased the number of mixtures to two and the separation is shown in Figure 39. As we continuously increased the

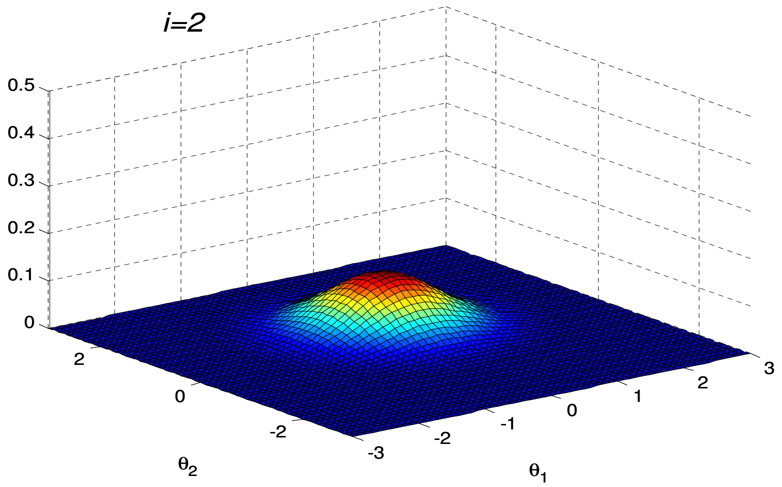


Figure 33. The lower bound estimate that touches the likelihood function with iteration 2 using a random initial estimate.

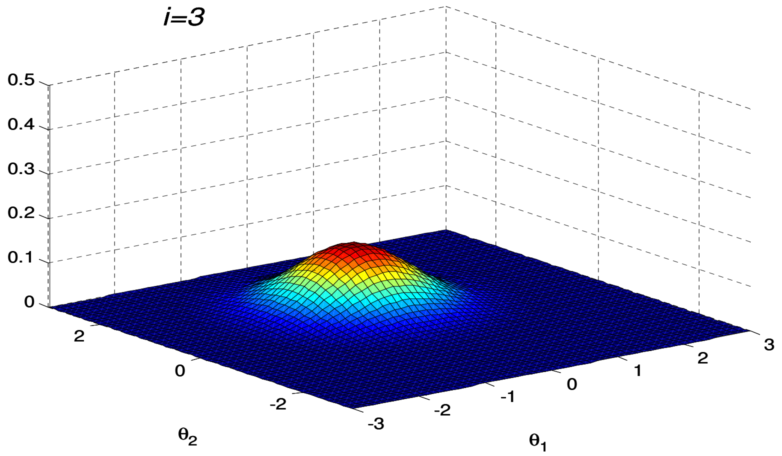


Figure 34. The lower bound estimate that touches the likelihood function with iteration 3 using a random initial estimate.

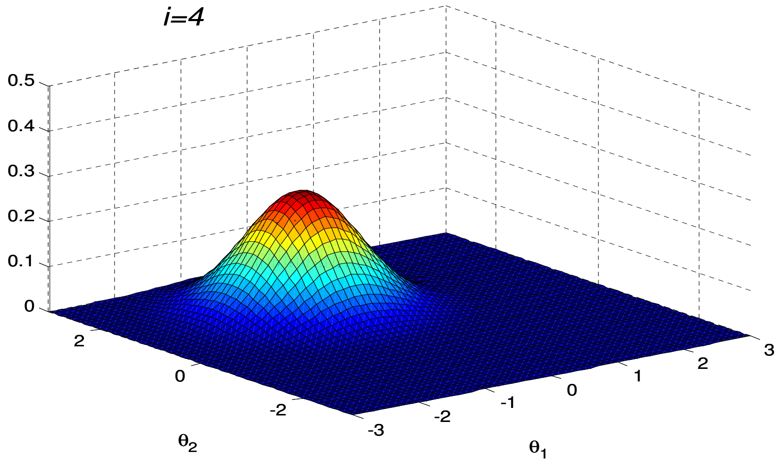


Figure 35. The lower bound estimate that touches the likelihood function with iteration 4 using a random initial estimate.

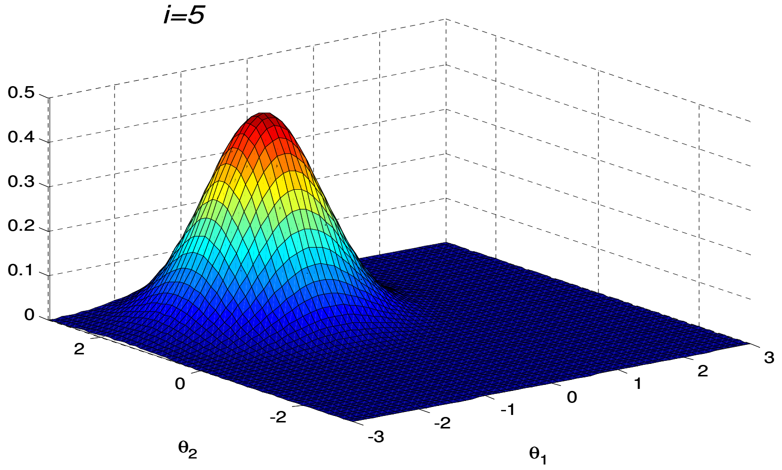


Figure 36. The lower bound estimate that touches the likelihood function with iteration 5 using a random initial estimate.

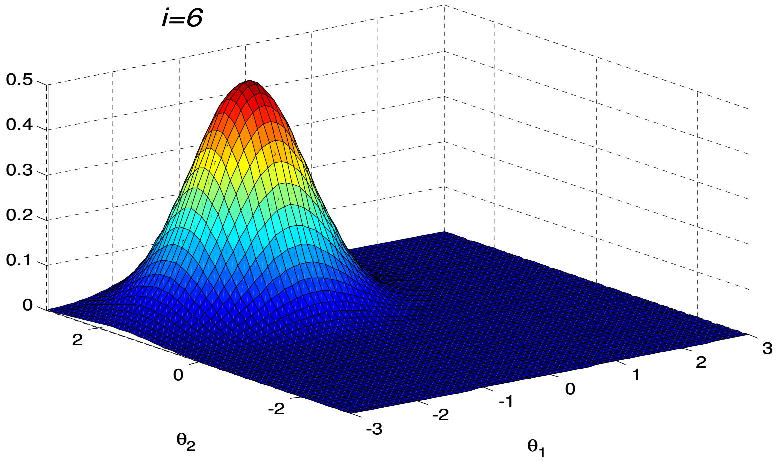


Figure 37. The lower bound estimate that touches the likelihood function with iteration 6 using a random initial estimate.

number of mixtures to three as shown in Figure 40, the boundary of separation between two classes improved further as compared to using fewer mixtures.

3.4 Maximum Likelihood or Bayes' Method

It is often the case where Maximum Likelihood and Bayes' are compared. Whether one is better than the other is more of a design issue which is dependent of the available resources at our disposal. In order to limit our descriptions of the advantages and disadvantages of the two, we will cite 2 important practical issues. The training data, and computational complexity.

Given an infinite data, both Maximum Likelihood and Bayes' method are equivalent in the asymptotic limit sense, assuming each has prior distributions that do not preclude the true solution. It should be apparent by now that we assume a parametric solution $p(\boldsymbol{x}|\widehat{\theta})$ for Maximum Likelihood. This condition however may not be true for Bayesian making it more general than ML. Its implication is that, Bayesian method has the characteristics of improving its performance with the addition of training points where Maximum Likelihood may render its solution to be unchanged [42]. With regards to the computational

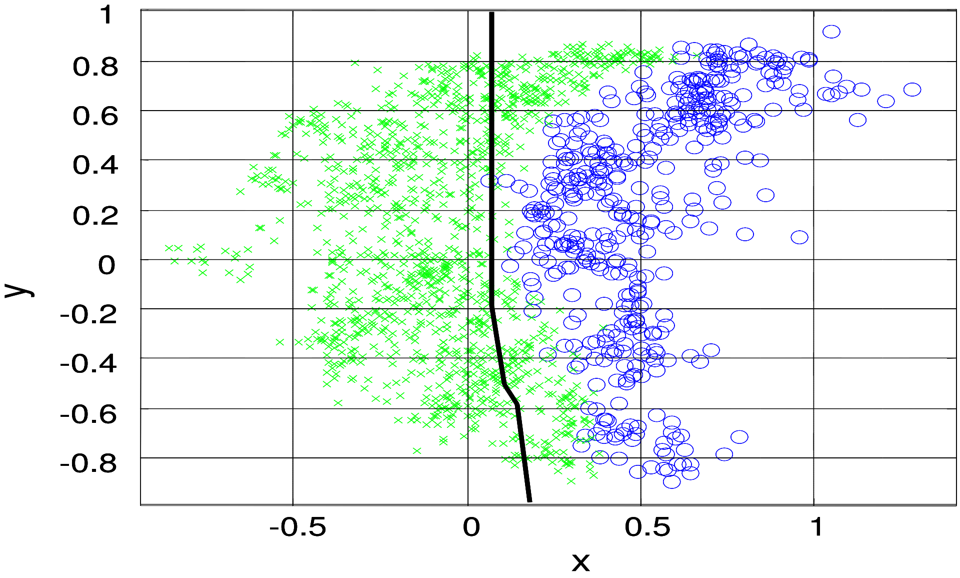


Figure 38. Boundary created by Expectation-Maximization using a single Gaussian mixture.

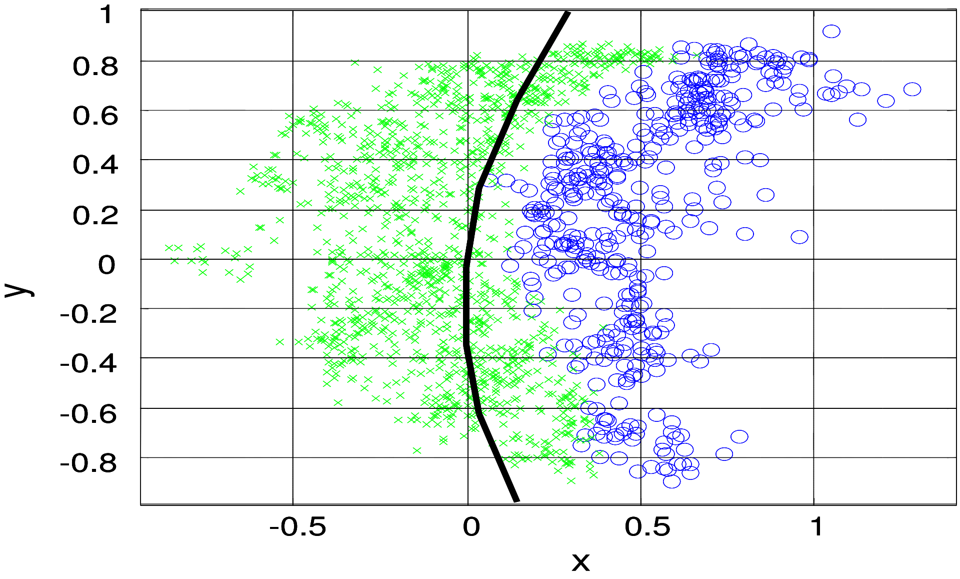


Figure 39. Boundary created by Expectation-Maximization using two Gaussian mixtures.

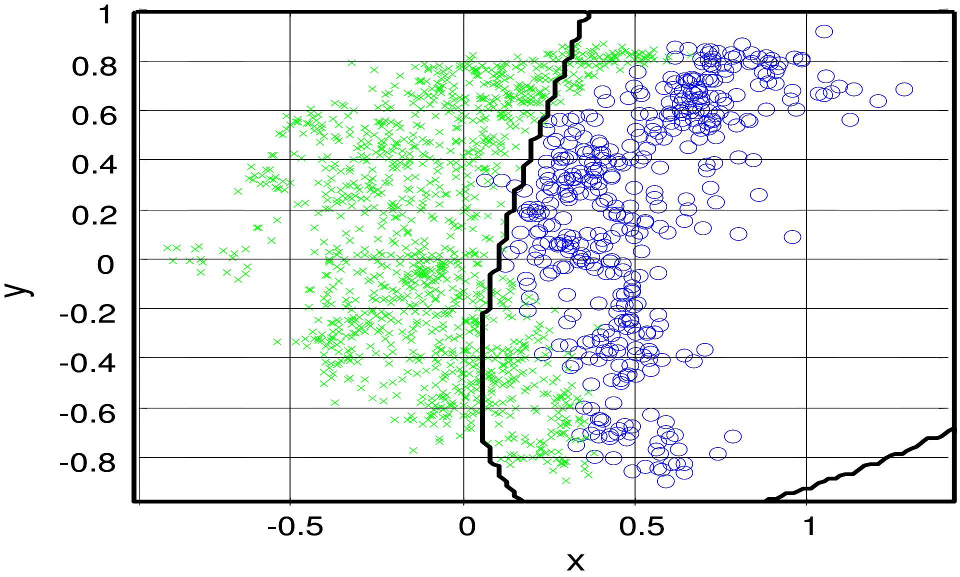


Figure 40. Boundary created by Expectation-Maximization using three Gaussian mixtures.

complexity issue, Maximum Likelihood requires only techniques used in calculus while the latter may involve multi-dimensional integration. Another issue related to computational complexity is the simplicity of the solution for the Maximum Likelihood where we arrive to a single estimate. Unlike in Bayesian we get more complicated answer in a form of weighted parameters. Lastly, as a direct consequence of computation complexity, since Maximum Likelihood requires a simpler solution, then it requires a shorter amount of time than the Bayesian approach.

3.5 Sufficient Statistics Background

We have shown the important role of MLE in parameter estimation in which the unknown parameter $\boldsymbol{\theta}$ is best estimated by maximizing the probability using the actual observed samples. Thus, an inference of a good parameter estimate is dependent with the actual observed data. In real applications, there could be a lot of parameters to be estimated using a very huge training data. This scenario would lead to a problem of directly computing Equation 23. It is desirable to find a feasible method in avoiding the difficulties of setting up the needed large parameters and at the same time, without compromising the estimate of $\boldsymbol{\theta}$. Depending on the distribution, computational feasible solutions that simplify the problem of directly computing Equation 23 is possible through Sufficient Statistics .

By virtue of Sufficient Statistics, we refer to a vector-valued function $\boldsymbol{s}$ of the sample in the training data D which is directly related in estimating the parameter $\boldsymbol{\theta}$. Furthermore, we define that $\boldsymbol{s}$ is *sufficient* for θ if it is shown that given $p(D|\boldsymbol{s}, \boldsymbol{\theta})$ is independent of θ. Thus, the *Factorization theorem*

$$p(D|\boldsymbol{\theta}) = g(\boldsymbol{s}, \boldsymbol{\theta})h(D) \ , \tag{33}$$

states that if the $p(D|\boldsymbol{\theta})$ can be written as a product where $g(.,.)$ is independent of the training data $h(D)$, then $\boldsymbol{s}$ is said to be the Sufficient Statistics of $\boldsymbol{\theta}$. Given this notion that Sufficient Statistics simplify the solution to calculating directly Equation 23, we can instead calculate for $\boldsymbol{s}$ rather than $\boldsymbol{\theta}$. Thus, it is very important to show that $\boldsymbol{s}$ would result to a good estimate or at least as good as MLE. In this part, we use the theory of Sufficient Statistics and compare its result to the MLE case. Again, for simplicity we consider the case of a Normal distribution with a known covariance and unknown mean. In deriving the Sufficient Statistics $\boldsymbol{s}$, the Factorization theorem is invoked and starting from the likelihood function

$$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp[-\frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\theta})] \ . \tag{34}$$

Since

$$\prod_{k=1}^{n} A \ exp[b] = A^n \ exp[\sum_{k=1}^{n} b], \tag{35}$$

we get

$$p(D|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{dn}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}} exp[-\frac{1}{2}\sum_{k=1}^{n}(\boldsymbol{\theta}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_k + \boldsymbol{x}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_k)] \ . \tag{36}$$

By rearranging and making sure that one factor is independent of $\boldsymbol{\theta}$, we get

$$p(D|\boldsymbol{\theta}) = exp[-\frac{n}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k)]$$
$$\cdot \frac{1}{(2\pi)^{\frac{dn}{2}}|\Sigma|^{\frac{n}{2}}} exp[-\frac{1}{2}\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_k]. \tag{37}$$

Considering the first factor which is a function of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ we have

$$g(\boldsymbol{s}, \boldsymbol{\theta}) = exp[-\frac{n}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k)] \ , \tag{38}$$

and from the equation above, it is obvious that the Sufficient Statistics $\boldsymbol{s}$ is given as

$$\boldsymbol{s} = \sum\nolimits_{k=1}^{n} \boldsymbol{x}_k \ , \tag{39}$$

Equation 38 can be rearranged as

$$g(\boldsymbol{s}, \boldsymbol{\theta}) = exp[-\frac{n}{2}(\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\frac{1}{n}\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k)] \ , \tag{40}$$

where the Sufficient Statistics becomes

$$\boldsymbol{s} = \frac{1}{n}\sum\nolimits_{k=1}^{n} \boldsymbol{x}_k \ , \tag{41}$$

which happens to be $\widehat{\boldsymbol{\theta}}$ particularly $\widehat{\mu}$ of the MLE given in Equation 32, thus

$$\boldsymbol{s} = \widehat{\boldsymbol{\mu}} \ . \tag{42}$$

Equation 42 is very important because we have established that $\boldsymbol{s}$ is equal to MLE's $\widehat{\boldsymbol{\theta}}$.

3.6 Applications of Sufficient Statistics

The notion of Sufficient Statistics is applicable to most of the exponential family which includes the Gaussian distribution. Thus, we can also apply this theory to speech recognition where we approximate speech as Gaussian. Moreover, there are distributions which this cannot be applied simply because Factorization theorem does not hold like the Cauchy distribution. It follows that as long as the

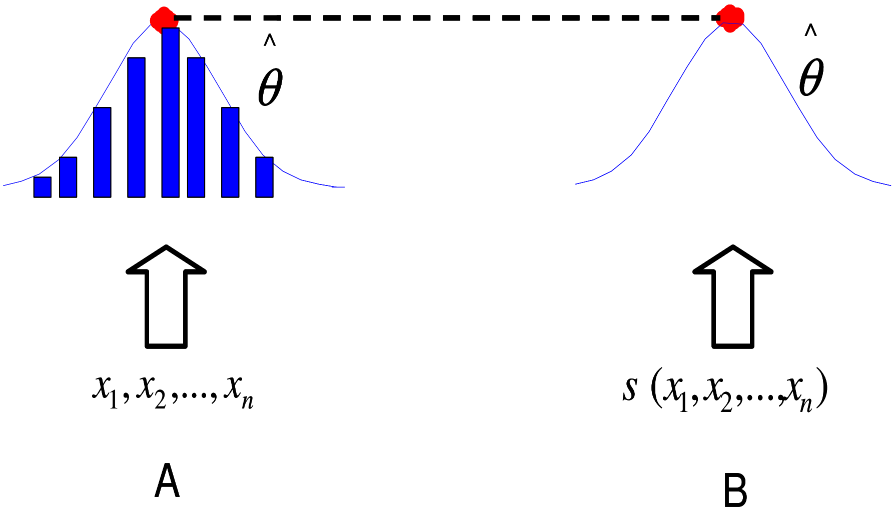


Figure 41. The differences between keeping the training samples and keeping only the Sufficient Statistics.

Sufficient Statistics exists, the sample mean and sample covariance are good estimators of the true mean and the true covariance. Sufficient Statistics can be interpreted as way of estimating the parameters without resorting to using the actual observed data itself. Since the basic foundation of Sufficient Statistics has already been discussed, the wide-range applications of HMM-Sufficient Statistics will be presented.

3.6.1 Data Reduction

Consider the independent and identically distributed random variables x_1,x_2,...,x_n having a distribution $\boldsymbol{Z}(\boldsymbol{\theta})$. Suppose that in Figure 41 given with two observation points A and B where in A one can have direct access of observing the entire data x_1,x_1,...,x_n and in B, one can only have an access to the Sufficient Statistics $s(x_1, x_2, ..., x_n)$. Between the two observation points, it is obvious that we can get more information such as the discrete distribution and its corresponding parameters $\widehat{\boldsymbol{\theta}}$ when observing at point A rather than at B. However, if one is only interested in finding $\widehat{\boldsymbol{\theta}}$ then point B gives us as much information as observing at point A. Most applications in pattern classification and in our adaptation approach, we are only interested of $\widehat{\boldsymbol{\theta}}$. The benefit in terms of data reduction should be clear by now. If one's application only requires $\widehat{\boldsymbol{\theta}}$ then it is practical to keep data in terms of Sufficient Statistics $s(x_1, x_2, ..., x_n)$ which is compact in size rather than keeping all of the observed data x_1,x_2,...,x_n.

3.6.2 Sufficient Statistics as a Better Estimator

Applications of Sufficient Statistics do not end in data reduction by keeping $\boldsymbol{s}$ and throwing away the rest of the data. Another important use of Sufficient Statistics is that it can serve as a basis for an estimator. The Rao-Blackwell Theorem states that given $\widehat{\boldsymbol{\theta}}$ be a finite-varianced estimator of θ and suppose that we have $\boldsymbol{T}$ which is the Sufficient Statistics $(\boldsymbol{s}_1, ..., \boldsymbol{s}_n)$ of $\boldsymbol{\theta}$. The new estimate $\widetilde{\boldsymbol{\theta}}$ using the Sufficient Statistics $\boldsymbol{T}$ is

$$\widetilde{\boldsymbol{\theta}} = \boldsymbol{E}(\widehat{\boldsymbol{\theta}}|\boldsymbol{T}) \ , \tag{43}$$

such that

$$\boldsymbol{T} = \boldsymbol{s}_1, ..., \boldsymbol{s}_n \ , \tag{44}$$

which results to a better estimator in MSE sense

$$\boldsymbol{E}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{2} \leq \boldsymbol{E}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{2}. \tag{45}$$

The equality is strict unless $\widehat{\boldsymbol{\theta}}$=$\widetilde{\boldsymbol{\theta}}$. Suppose, that $\boldsymbol{s}_1, ..., \boldsymbol{s}_n \sim \boldsymbol{Z}(\boldsymbol{\theta})$ and we estimate $\boldsymbol{\theta}$. By starting with the unbiased estimator $\boldsymbol{\theta} = \boldsymbol{s}_1$, Rao-Blackwellization results to

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \boldsymbol{E}[\boldsymbol{s}_1|\boldsymbol{T} = t] \\ &= \frac{1}{n}\boldsymbol{E}[\boldsymbol{s}_1| = \sum\nolimits_{k=1}^{n} \boldsymbol{s}_k = t], \end{aligned} \tag{46}$$

since the Sufficient Statistics $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ are IID, then

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \boldsymbol{E}[\boldsymbol{s}_1|\boldsymbol{s}_1 + \boldsymbol{s}_2 + ... + \boldsymbol{s}_n] \\ &= \frac{\boldsymbol{s}_1 + \boldsymbol{s}_2 + ... + \boldsymbol{s}_n}{n}. \end{aligned} \tag{47}$$

The intuition of Equation 47 is that, Sufficient Statistics is used as a new estimator which is actually a better estimator in the MSE sense. Its implication is of great value since it paves the way of breaking down extensive computations in solving equation 24. Consider that you want to estimate the parameter $\boldsymbol{\theta}$ using ML with a huge training data. With the relationship of ML, Sufficient Statistics, and by using Rao-Blackwellization, we can segregate the training data into several parts and calculate the corresponding Sufficient Statistics of each part. If we take the average of these kept Sufficient Statistics in Equation 47 we can estimate our new $\widehat{\boldsymbol{\theta}}$. This results to a reduction of computation time as one can possibly distribute computation loads to different processors without affecting the final estimate. This is not just limited to parallel processing but works in selective training also makes use of this technique [43]. Moreover, the proposed rapid adaptation makes use of this as well, which will be discussed in later chapters.

3.7 Summary

This chapter is dedicated in explaining the general concept of Sufficient Statistics which is used in the proposed rapid adaptation. To understand this, we started with the basic pattern classification concept which include Bayes' decision rule, Bayesian training and Maximum Likelihood Estimation. The relationship between the latter and Sufficient Statistics is also established. Moreover, the applications of Sufficient Statistics which includes parameterizing the data into Sufficient Statistics for data reduction and parallel training by virtue of Rao-Blackwellization are presented.

4. Hidden Markov Models-Sufficient Statistics and N-best Speakers Selection

Adaptation time and adaptation data are two of the most common issues in improving the performance of speech recognition. The proposed approach addresses these two problems. First, through the use of HMM-Sufficient Statistics, adaptation algorithms become more efficient and faster as adaptation data are processed in advance and only the statistical parameters are kept. Secondly, using the concept of N-best speaker selection, adaptation is possible without collecting adaptation utterances from the user.

4.1 The Expectation-step

In this section, we will expand the general concept of Sufficient Statistics discussed in the previous chapter, and in our application we refer to HMM-Sufficient Statistics. Solving for the HMM-Sufficient Statistics, is the first step in realizing the rapid adaptation approach. This constitutes the Expectation-step (E-step) which is analogous to computing the lower bound in Figure 27, in which we refer as the Q function $\boldsymbol{Q}^t(\theta)$. It basically requires the calculation of the probability of the observation sequence $\boldsymbol{O}$ given the model λ i.e., $P(\boldsymbol{O}|\lambda)$. Since it is not feasible to calculate this directly, a more efficient way is to use Forward procedure to calculate the required probabilities. A similar approach which is the time-reversed version of Forward procedure known as the Backward procedure is available. By using these two algorithms, the HMM-Sufficient Statistics can be computed.

4.1.1 The Forward and Backward Procedure

These algorithms evolve through the fact that only the likelihood of generating an observation and the likelihood of being in a particular state at a given time is important. The forward procedure computes the required probabilities $\alpha_j(t)$ in a trellis, as the HMM unfolds through time. The subscript j and t denotes the state $1 \leq j \leq N$ and the time $1 \leq t \leq T$ respectively. The forward variable $\alpha_j(t)$ is defined as

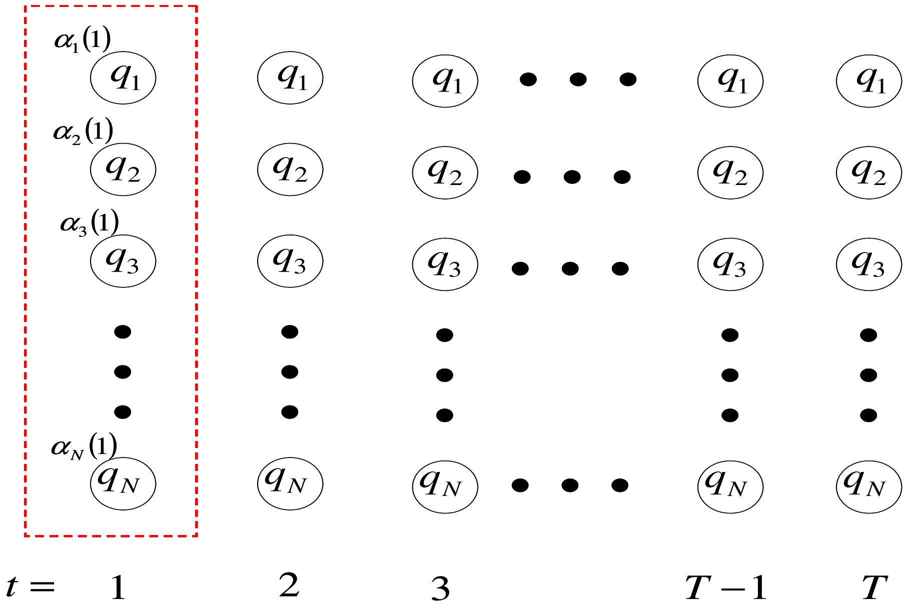


Figure 42. Initialization of the forward probabilities

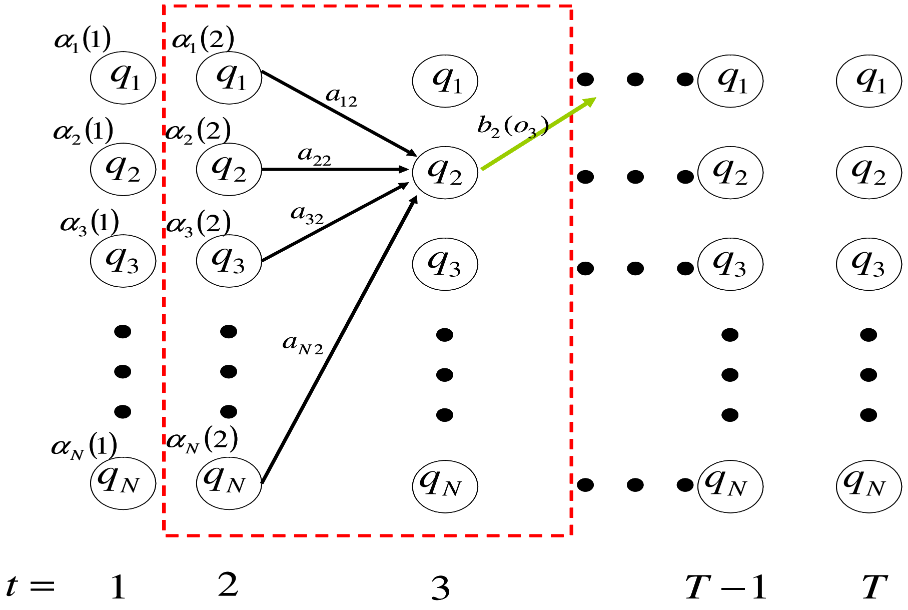


Figure 43. Computation of the forward probabilities by induction

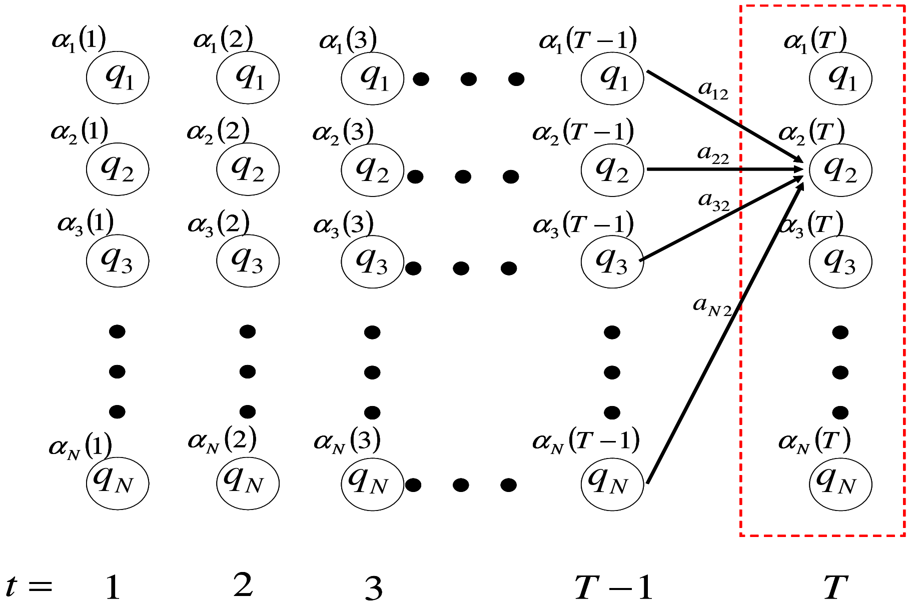


Figure 44. The last phase of Forward algorithm which result to the terminal forward probabilities

$$\alpha_j(t) = P(\boldsymbol{o}_1\boldsymbol{o}_2 \ldots \boldsymbol{o}_t, q_t = j|\lambda). \tag{48}$$

The algorithm is primarily composed of three parts. The first part is the initialization phase shown in Figure 42. This is the start of the computation of the forward probability $\alpha_j(t)$ known as the initial conditions given by

$$\alpha_j(1) = a_{1j}b_j(\boldsymbol{o}_1), \tag{49}$$

where b_j denotes the forward probability of the emission of the visible state and $\boldsymbol{o}_1$ is the observation at $t = 1$. After the initialization, the forward probabilities shown in Figure 43 are then computed by induction through forward recursion

$$\alpha_j(t) = [\sum\nolimits_{i=1}^{N} \alpha_i(t-1)a_{ij}]b_j(\boldsymbol{o}_t). \tag{50}$$

The final phase shown in Figure 44 is where the calculation of the forward probabilities terminate at t=T for $1 \leq j \leq N$

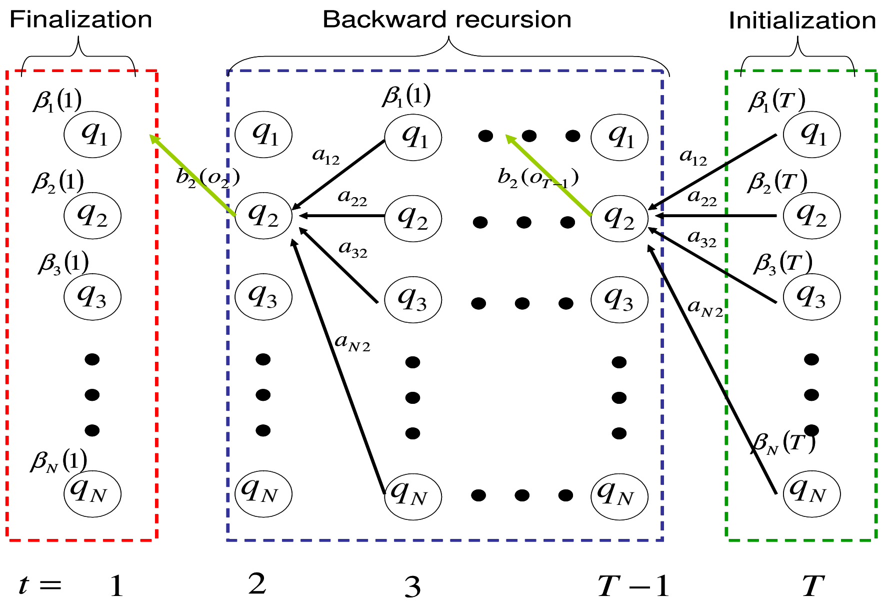


Figure 45. The Backward procedure showing the initialization phase, backward recursion and the finalization phase.

$$\alpha_j(T) = \sum\nolimits_{i=1}^{N} \alpha_i(T) a_{ij}. \tag{51}$$

From the final phase, it is obvious that $P(\boldsymbol{O}|\lambda)$ can be computed by summing all of the terminal forward variables $\alpha_j(T)$

$$P(\mathrm{O}|\lambda) = \sum\nolimits_{j=1}^{N} \alpha_j(T). \tag{52}$$

The Backward procedure on the other hand is the time-reversed version of the Forward algorithm which allows for the calculation of the backward variables $\beta_i(t)$ for $1 \leq i \leq N$ and $1 \leq t \leq T$ and defined as

$$\beta_i(t) = P(\boldsymbol{o}_{t+1}\boldsymbol{o}_{t+2} \ldots \boldsymbol{o}_T | q_t = i, \lambda). \tag{53}$$

The trellis for the Backward procedure is shown in Figure 45 where the initialization takes place at t=T for $1 \leq j \leq N$ by setting

$$\beta_i(t) = a_{ij}, \tag{54}$$

the backward recursion allows to traverse the trellis, and is given by

$$\beta_i(t) = \sum\nolimits_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_t + 1)\beta_j(t+1). \tag{55}$$

Lastly, the final computed probabilities for $1 \leq i \leq N$ is to set t=1 in Equation 55.

4.1.2 HMM-Sufficient Statistics Parameters

As a direct consequence of the Forward and Backward procedures, we can calculate the parameters of the HMM-Sufficient Statistics

$$\boldsymbol{m}_{im}^{spkr} = \sum\nolimits_{r=1}^{R_{spkr}} \sum\nolimits_{t=1}^{T_r} L_{im}^{r}(t)\boldsymbol{O}_t^r \ , \tag{56}$$

$$\boldsymbol{v}_{im}^{spkr} = \sum\nolimits_{r=1}^{R_{spkr}} \sum\nolimits_{t=1}^{T_r} L_{im}^{r}(t)\boldsymbol{O}_t^r \boldsymbol{O}_t^{rT} \ , \tag{57}$$

where L_{im}^r is the probability of mixture component occupancy while $\boldsymbol{m}_{im}^{spkr}$ and $\boldsymbol{v}_{im}^{spkr}$ are the mean and variance of a particular state i and mixture component m respectively as represented by the subscript im. The observation vector is denoted by $\boldsymbol{O}$. The index *spkr* refers to the particular speaker where the training data comes from, and will be discussed further in the N-best speaker section.

Aside from Equations 56 and 57, we also calculate for the accumulated probability of the mixture occupancy L_{im}^{spkr} given as

$$L_{im}^{spkr} = \sum\nolimits_{r=1}^{R_{spkr}} \sum\nolimits_{t=1}^{T_r} L_{im}^{r}(t) \ , \tag{58}$$

where R_{spkr} is the total number of speakers in the training data. Moreover, the state transition occupancy is calculated as

$$L_{ij}^{spkr} = \sum\nolimits_{r=1}^{R_{spkr}} \sum\nolimits_{t=1}^{T_{r-1}} L_{ij}^{r}(t). \tag{59}$$

The parameters given by Equations 58 and 59, which are computed during the E-step are also kept together with Equations 56 and 57. The proposed method is specifically designed to operate using HMM-Sufficient Statistics parameters instead of the actual observations provided by the actual samples of the adaptation data. These parameters are then used for a rapid model adaptation to be discussed in Chapters 5 and 6.

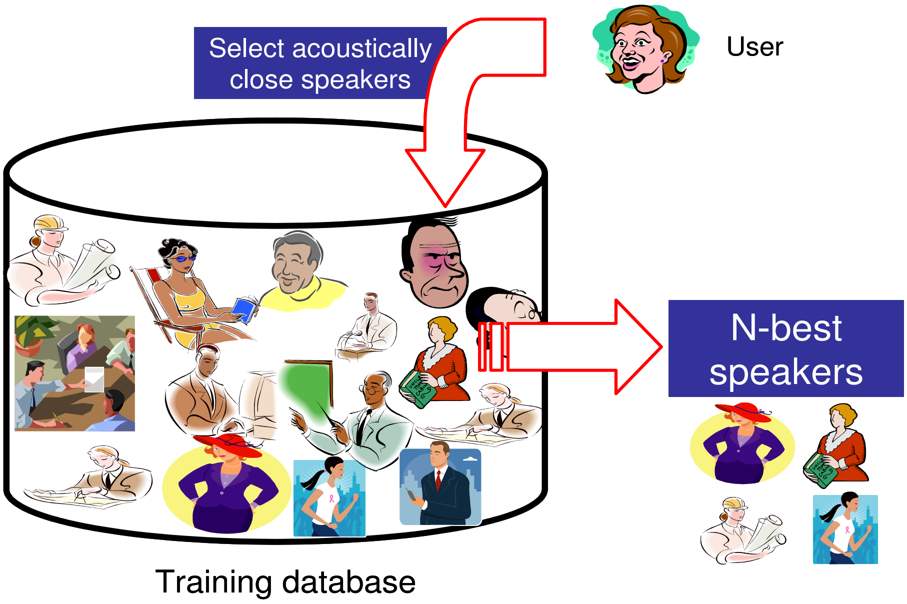


Figure 46. Selecting adaptation data from the training database that are acoustically close to the user using a single utterance

4.2 N-best Speakers' HMM-Sufficient Statistics as Adaptation Data

In the previous section, we discussed about keeping the HMM-Sufficient Statistics instead of the actual samples of the adaptation data $\boldsymbol{o}$. In this section, we will discuss another advantage of the proposed method, which is the ability to use the training database as adaptation data. Instead of collecting the actual adaptation data from the user, we can bypass this by using the N-best speakers in the training database as the adaptation data. This approach is done concurrently with the presumption, that given a multiple training databases with a huge spectrum of speakers of different genders and age-groups, there exists N-nearest speakers (N-best) that are acoustically similar to the arbitrary utterance which can be used as adaptation data, as shown in Figure 46. With this approach, the time needed in gathering the actual adaptation data of the test speaker is bypassed. By combining the concept of HMM-Sufficient Statistics and the N-best speaker selection, the training database is converted to HMM-Sufficient Statistics. This

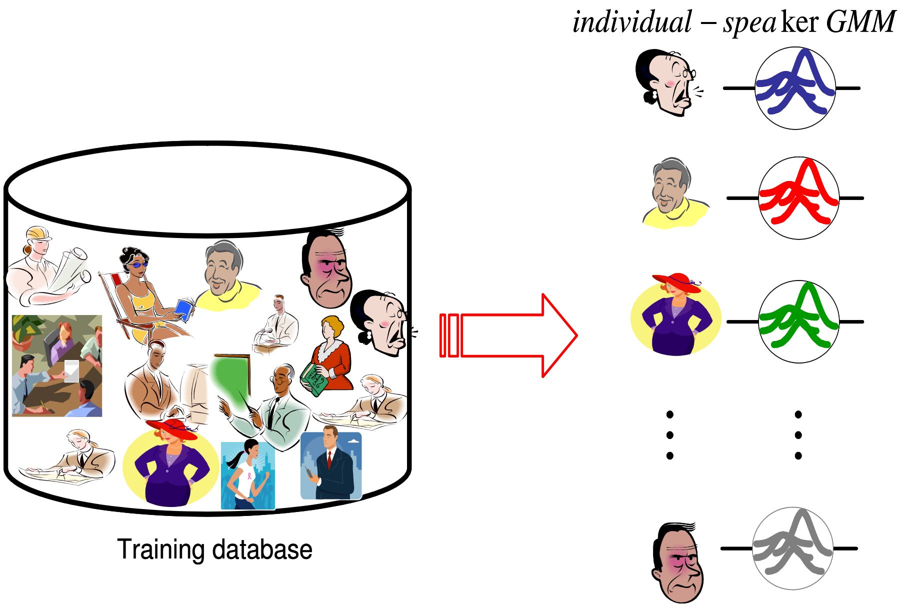


Figure 47. Creating individual-speaker GMMs from the training database

means that each of the speaker in the training database has a corresponding HMM-Sufficient Statistics which will be used when adaptation is carried out. Prior to the selection of the speakers' HMM-Sufficient Statistics, the system needs to identify the N-best speakers, and this process is done by creating individual-speaker Gaussian Mixture Models (GMMs) as shown in Figure 64, each of the speaker in the training database has a corresponding GMM and these are trained offline.

When all of the individual-speaker GMMs are prepared, N-best speaker selection is done online as shown in Figure 48 and described as follows:

1) The noisy utterance is denoised using SS prior to GMM selection followed by the parameterization to MFCC. Since there is a trace of residual noise after SS, low-power mfcc frames in the denoised utterance are removed and only the high-power MFCC frames are retained. In this way, the effects of the residual noise that is present in the silence or unvoiced region is reduced. This however does not affect the overall performance of the system since our models are individual-speaker GMMs.

2) We use the MFCC as observation vectors to find the likelihood given the

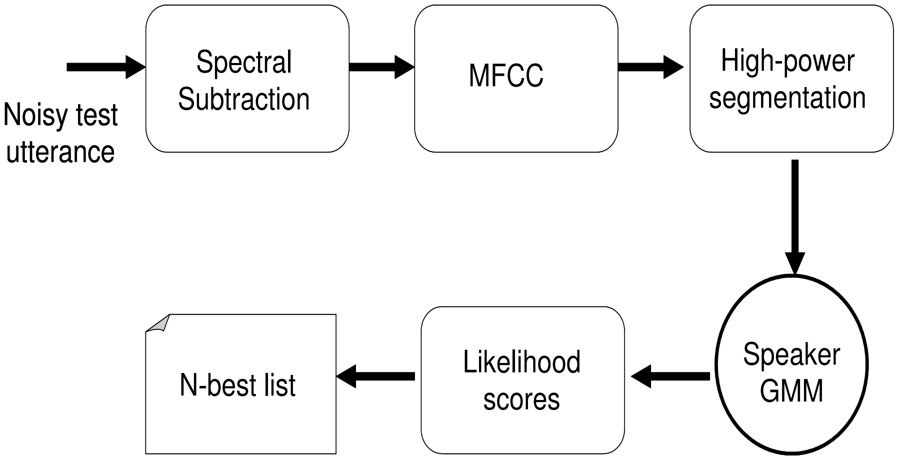


Figure 48. GMM speaker selection using the noisy utterance.

individual-speaker GMMs which results to likelihood scores.

3) From the likelihood scores, only N-best speakers that are close to the test utterance are selected. Meaning, only speakers in the GMMs that are acoustically close to the test utterance which gives the N-highest likelihood scores will be generated in the N-best speakers list.

4) From the N-best speakers' list, the system will automatically select the corresponding HMM-Sufficient Statistics of each speaker during the actual adaptation to be discussed in Chapters 5 and 6.

The advantage of N-best speaker selection is the ability to search for the most suitable adaptation data in the training database using the likelihood criterion. Given a huge pool of speakers in the database, not all of these are close to the test utterance. Although an accurately trained model can be achieved when using all of the speakers in the training database, the trained model has no good speaker discrimination property. By using N-best speaker selection, we improve speaker discrimination by choosing only the adaptation data from the neighborhood of N-best speakers. In effect, recognition performance improves as compared to using all of the adaptation data where the recognition performance is generally flat. Figures 49 and 50 is the plot of the actual distribution of the male and female speakers respectively in our database. Due to the high dimensionality of the Gaussian models used to describe each of the speaker, we use Principal Component analysis in projecting to two dimensions. In these figures, we illustrate the peaking of the recognition performance corresponding to the N-best neighborhood. Also, the flat recognition performance is due to the over-averaging of Gaussians when using indiscriminately all of the data.

4.3 Summary

This chapter expands the general concept of Sufficient Statistics that was loosely described in Chapter 3. Discussions evolved primarily on HMM-Sufficient Statistics, including the algorithm needed in its calculation. We have also discussed the mechanism on how to substitute actual users' adaptation data to using the training database by means of N-best speaker selection. With the concept of HMM-Sufficient Statistics and the mechanism to select adaptation data using N-best speakers, adaptation data is effectively processed beforehand and kept prior to actual adaptation.

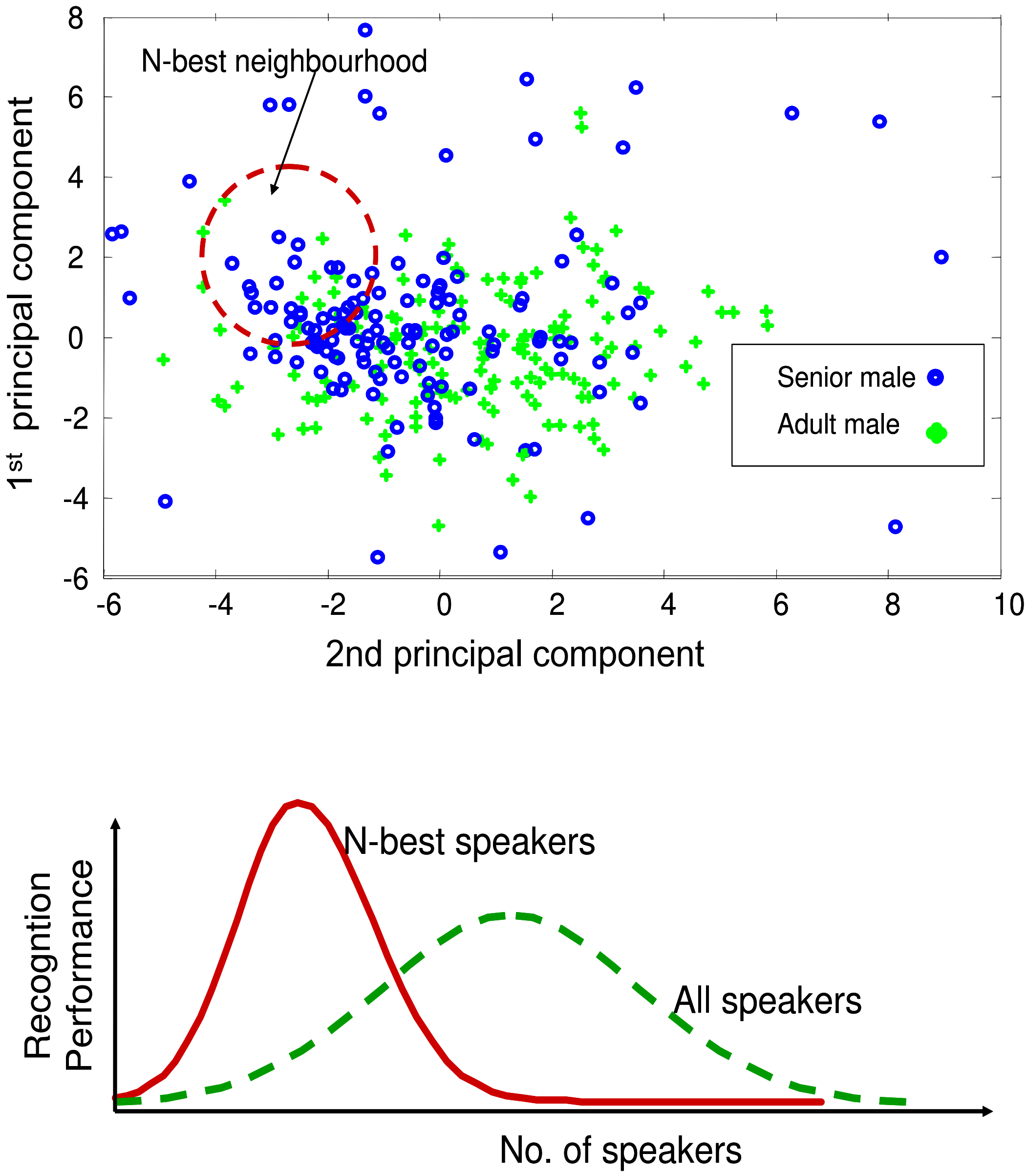


Figure 49. Optimizing recognition performance using N-best speakers selection (adult and senior male database).

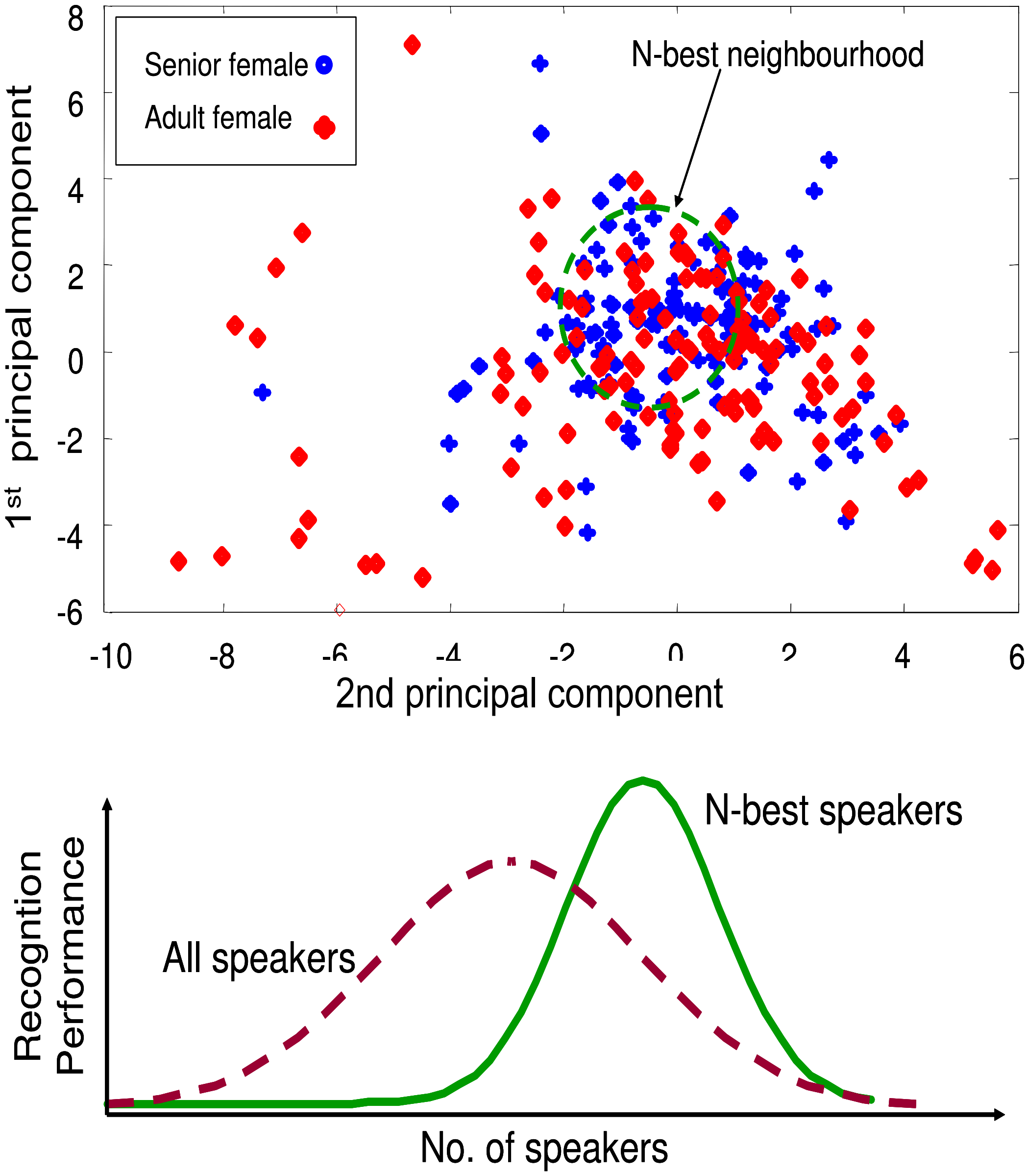


Figure 50. Optimizing recognition performance using N-best speakers selection (adult and senior female database).

5. Rapid Baum-Welch Unsupervised Speaker Adaptation based on N-best Speakers' HMM-Sufficient Statistics

In the previous chapter we discussed about the importance of the Q function and the probabilities that are attached to it. As a result we employ the E-step and calculate the HMM-Sufficient Statistics. In this chapter, we aim for the maximization of the Q function which is basically the M-step. This process completes the Baum-Welch re-estimation which results to an updated model where $P(\boldsymbol{O}|\lambda)$ is maximized. The conventional way of updating the model parameters using Baum-Welch is to execute the whole E-M process as follows

$$C_{im}^{adp} = \frac{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)}{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{i}^{r}(t)}, \tag{60}$$

$$\boldsymbol{\mu}_{im}^{adp} = \frac{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)\boldsymbol{o}_t^r}{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)}, \tag{61}$$

$$\boldsymbol{\Sigma}_{im}^{adp} = \frac{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)(\boldsymbol{o}_t^r - \boldsymbol{\mu}_{im}^{adp})(\boldsymbol{o}_t^r - \boldsymbol{\mu}_{im}^{adp})^T}{\sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)}, \tag{62}$$

and

$$a_{ij}^{adp} = \frac{\sum_{t=1}^{T_r} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t)\beta_t(j)}{\sum_{t=1}^{T_r} \alpha_{t-1}(i)\beta_{t-1}(i)} \tag{63}$$

where C_{im}^{adp}, $\boldsymbol{\mu}_{im}^{adp}$ $\boldsymbol{\Sigma}_{im}^{adp}$, and a_{ij}^{adp} are the updated mixture weight, mean, covariance matrix and updated transition probability respectively. Moreover, Equations 61 and 62 are dependent of the actual observation data $\boldsymbol{o}$.

In realizing a rapid Baum-Welch model adaptation, we use HMM-Sufficient Statistics as adaptation data, thus performing beforehand the E-step as discussed in Chapter 4.1 and keeping the parameters $\boldsymbol{m}_{im}^{spkr}$, $\boldsymbol{v}_{im}^{spkr}$, L_{im}^{spkr}, and L_{ij}^{spkr} given in Equations 57-59. During the actual adaptation process, N-best speaker selection is used to select the speakers' HMM-Sufficient Statistics and use these parameters in Baum-Welch model updating. The updated parameters using HMM-Sufficient Statistics are

$$C_{im}^{adp} = \frac{\sum_{s=1}^{S} L_{im}^{s}}{\sum_{s=1}^{S} \sum_{m=1}^{M} L_{im}^{s}} , \tag{64}$$

$$\boldsymbol{\mu}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}} , \tag{65}$$

$$\boldsymbol{\Sigma}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}} - \boldsymbol{\mu}_{im}^{adp} \boldsymbol{\mu}_{im}^{adp^T} , \tag{66}$$

and

$$a_{ij}^{adp} = \frac{\sum_{s=1}^{S} L_{ij}^{s}}{\sum_{s=1}^{S} \sum_{j=1}^{J} L_{ij}^{s}} \tag{67}$$

where C_{im}^{adp}, $\boldsymbol{\mu}_{im}^{adp}$, $\boldsymbol{\Sigma}_{im}^{adp}$, and a_{ij}^{adp} are the updated mixture weight, mean, covariance matrix and updated transition probability respectively. L_{im}^{s}, L_{ij}^{s}, $\boldsymbol{m}_{im}^{s}$, $\boldsymbol{v}_{im}^{s}$ are the probability of mixture component occupancy, the accumulated probability of the state occupancy, means and variance respectively of the selected N-best speakers. Since the HMM-Sufficient Statistics are already pre-computed offline, the actual online adaptation only requires performing the M-step which is transformed to simply summing and averaging HMM-Sufficient Statistics in which extensive computations are done offline. As opposed to Equations 61 and 62, the adapted mean and covariance given in Equations 65 and 66 are independent of the actual observation vector $\boldsymbol{o}$.

In the following sections, we will discuss the different implementation in the context of Baum-Welch using HMM-Sufficient Statistics. These include the single-template approach, multi-template, weighting of the HMM-Sufficient Statistics, linear interpolation and clustering of speakers. No matter what approach it takes, the Baum-Welch adaptation using HMM-Sufficient Statistics is mainly composed of two parts, the *offline* where speakers' HMM-Sufficient Statistics are

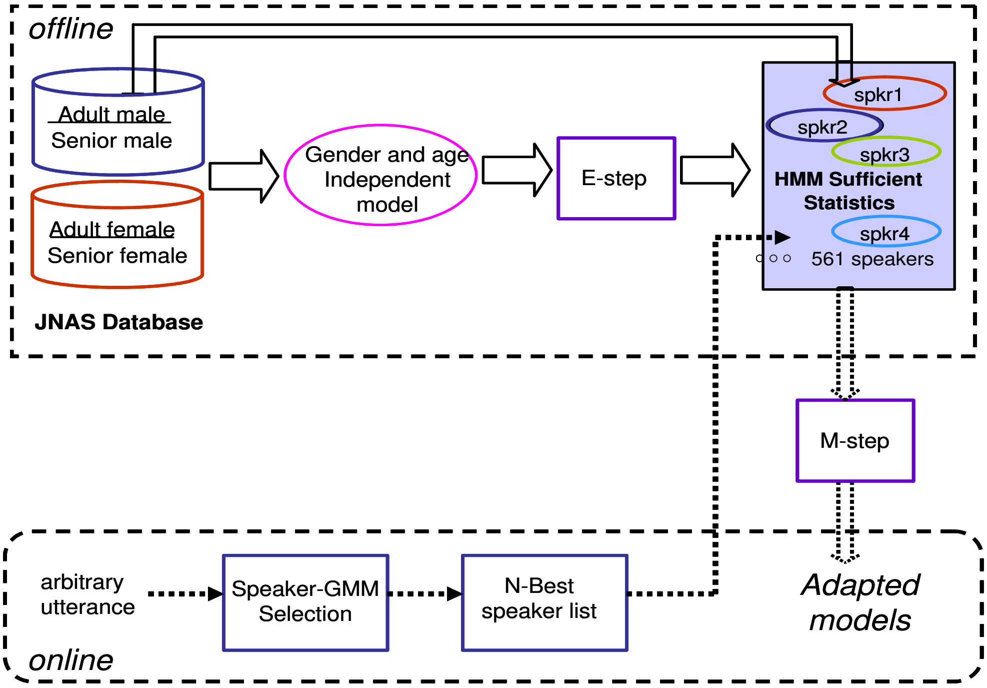


Figure 51. Block diagram of the conventional HMM-Sufficient Statistics adaptation.

created, and the *online* where speaker selection and the actual adaptation takes place.

5.1 Single-template HMM-Sufficient Statistics

Single-template HMM-Sufficient Statistics adaptation is shown in Figure 51. In the *offline* part, SI model is trained regardless of classes using all of the training data from the JNAS Adult and Senior database. Using this SI model, the E-step is performed which leads to the estimate of the HMM-Sufficient Statistics per speaker in the database, and these parameters are kept. In the *online* part, N-best speakers selection takes place which is immediately followed by the execution of the M-step which constitutes the actual adaptation and completes the whole process of Baum-Welch through updating of the target speaker's model parameters using the pre-calculated HMM-Sufficient Statistics.

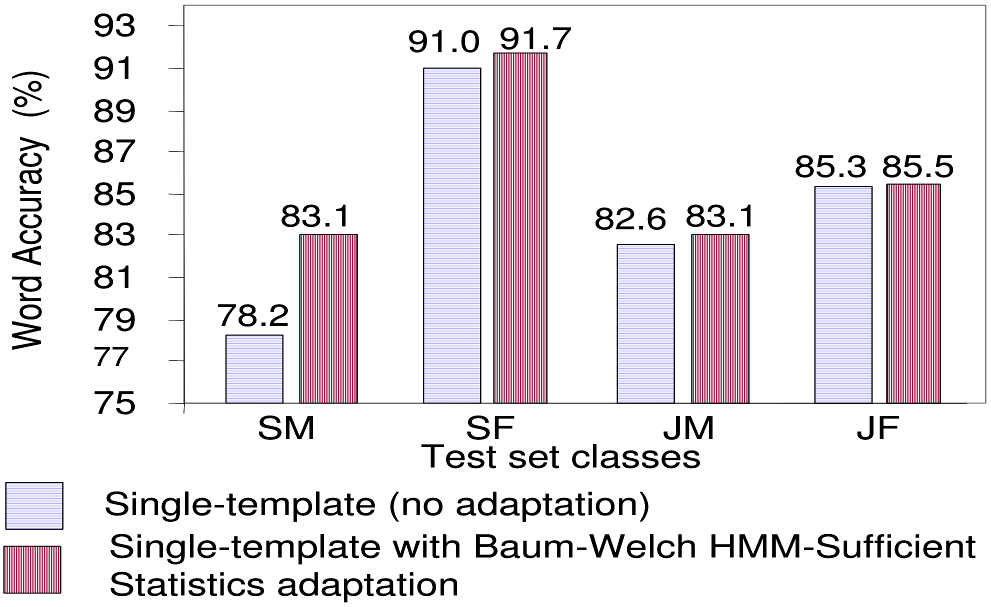


Figure 52. Recognition performance using single-template HMM-Sufficient Statistics adaptation with one iteration of Baum-Welch reestimation.

5.1.1 Results

Figure 52 shows the result of the single-template HMM-Sufficient Statistics adaptation based on one-iteration of Baum-Welch. In this graph, the recognition performance using only the SI model without adaptation is used as the baseline and the four classes of test sets used. The graph shows that the proposed method using a single arbitrary utterance of data performs better than using only the SI model without adaptation. Moreover, The Senior male (SM) which has an absolute improvement of 4.9% has the most significant improvement among all of the test classes.

5.1.2 Limitations of the Single-template HMM-Sufficient Statistics Adaptation

It is possible that the acoustical characteristics of the test speaker and the SI model is very dissimilar. Usually, Baum-Welch is iterated several times resulting to a progressively improved model estimate as discussed in Chapter 3.3. The

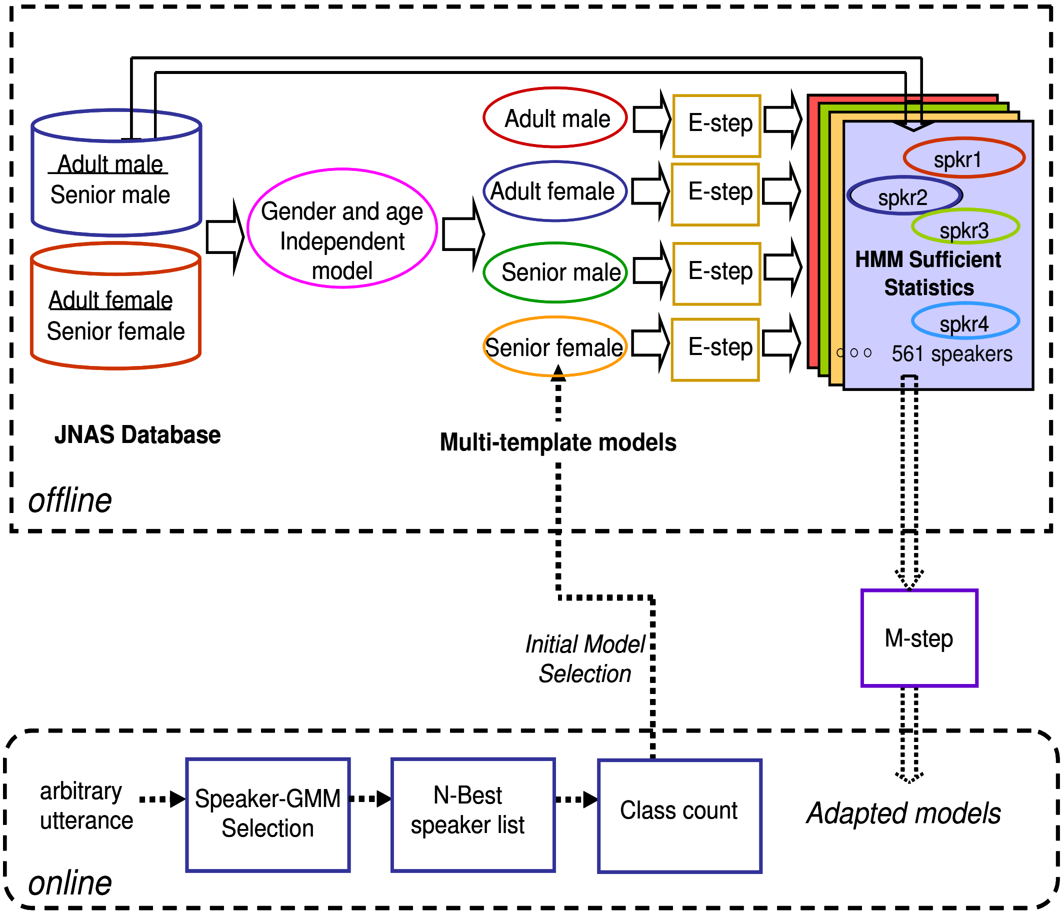


Figure 53. Block diagram of the multi-template HMM-Sufficient Statistics adaptation.

iterative process eliminates the mismatch due to several model updates. However, in the rapid Baum-Welch adaptation using HMM-Sufficient Statistics, we only use one-iteration of Baum-Welch owing to adaptation time constraint. Thus, model reestimation is not optimized. To counter the effect of the one-iteration constraint, a good initial model is needed, the one that is acoustically close to the test speaker.

5.2 Multi-Template HMM-Sufficient Statistics

To improve the performance of the one-iteration Baum-Welch, we introduce multi-template models. This approach resembles that of cluster based modelling [18] but more effective since we do not just choose an optimal model but we also incorporate adaptation. Figure 53 is the block diagram of the multi-template rapid Baum-Welch reestimation. From the SI model, multi-template HMMs are created namely: Adult male, Adult female, Senior male and Senior female. Consequently, by implementing the E-step, four sets of HMM-Sufficient Statistics for each speaker are created. Thus, gender and age information are emphasized and embedded in the HMM-Sufficient Statistics. In the *online* phase, N-best speakers are selected and followed immediately by the implementation of the M-step. As a result, the adapted model has an improved discrimination performance among different classes of speaker's acoustical characteristics. This method gives the system more degrees of freedom to choose the appropriate template model which is closer to the test utterance. This compensates the effect of using only a single iteration of the Baum-Welch approach. The merit of this is best explained in Figure 54. Selecting a template model close to the test utterance gives a better initial estimate of the lower bound than using a single-template model which is very essential when using only a single-iteration of Baum-Welch. The adapted parameters are as follows

$$C_{im}^{adp(temp)} = \frac{\sum_{s=1}^{S} L_{im}^{s(temp)}}{\sum_{s=1}^{S} \sum_{m=1}^{M} L_{im}^{s(temp)}} , \tag{68}$$

$$\boldsymbol{\mu}_{im}^{adp(temp)} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s(temp)}}{\sum_{s=1}^{S} L_{im}^{s(temp)}} , \tag{69}$$

$$\boldsymbol{\Sigma}_{im}^{adp(temp)} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s(temp)}}{\sum_{s=1}^{S} L_{im}^{s(temp)}} - \boldsymbol{\mu}_{im}^{adp(temp)} \boldsymbol{\mu}_{im}^{adp(temp)^T} , \tag{70}$$

and

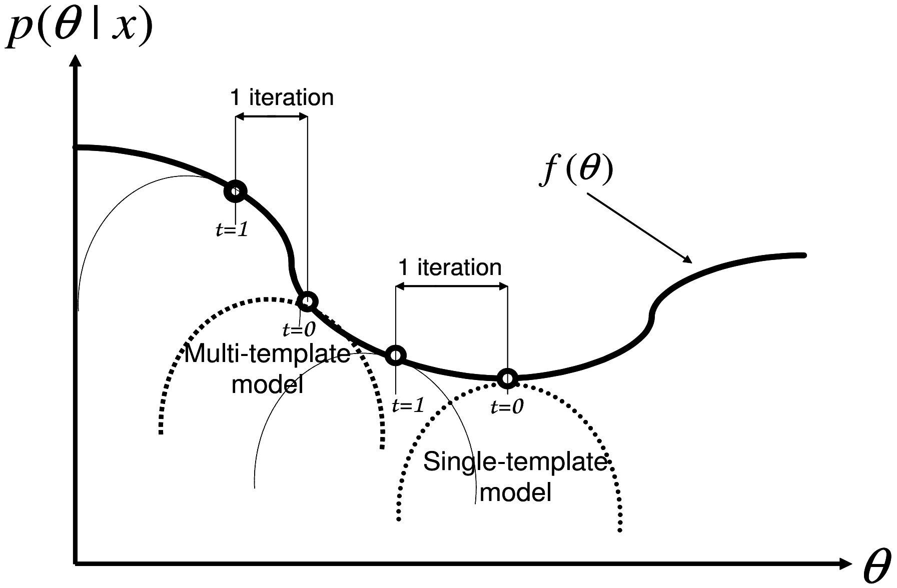


Figure 54. Template selection using multi-template models gives better estimate than using single-template model.

$$a_{ij}^{adp(temp)} = \frac{\sum_{s=1}^{S} L_{ij}^{s(temp)}}{\sum_{s=1}^{S} \sum_{j=1}^{J} L_{ij}^{s(temp)}} \tag{71}$$

where *temp* refers to the classes in the multi-template models (Adult male, Adult female, Senior male, and Senior female)

5.2.1 Results

In Figure 55 we compared the recognition performances for the three conditions. Using the single-template rapid Baum-Welch adaptation, multi-template models without adaptation, and when combining multi-template models with Baum-Welch adaptation. In this graph, it is apparent that the use of multi-template models together with the rapid Baum-Welch adaptation performs better than the single-template Baum-Welch and the unadapted multi-template models. This proves that the system's performance responds positively to the improved initial model using the multi-template models through model selection where the Baum-

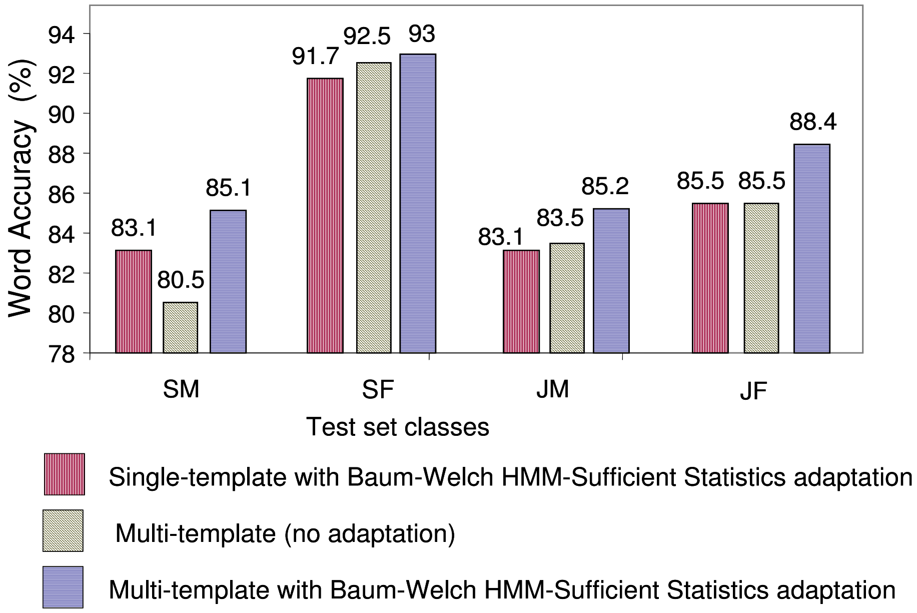


Figure 55. Recognition performance using multi-template HMM-Sufficient Statistics adaptation with one iteration of Baum-Welch reestimation.

Welch rapid adaptation can possibly select the models and Sufficient Statistics that are acoustically close to the test utterance.

5.2.2 Limitations of the Multi-template HMM-Sufficient Statistics Adaptation

The recognition performance and adaptation speed of this approach are dependent on the number of N-best speakers, S. Experiments showed that the optimal N-best is $S_{optimal} = 40$ which corresponds to a 10-second adaptation time [28] [44] [45]. If S is further reduced such that $S < S_{optimal}$, adaptation time is reduced with a trade-off of the recognition performance as illustrated in Figure 56. This is attributed to the fact that further decreasing S would result to insufficient data necessary to robustly estimate the target speaker's HMMs.

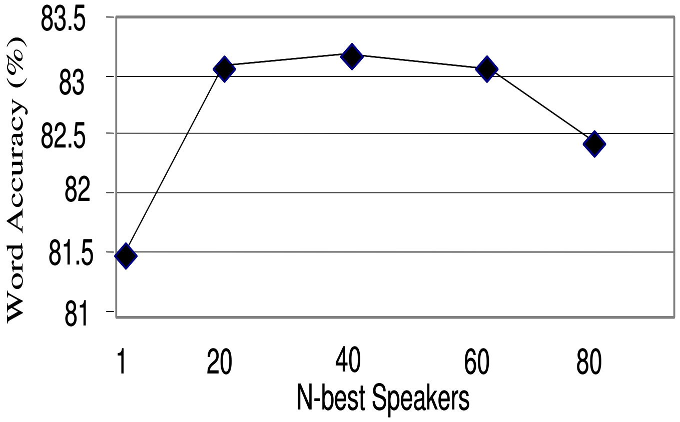


Figure 56. Relationship between N-best and recognition performance.

5.3 Weighting of the HMM-Sufficient Statistics

The single-template and multi-template Baum-Welch reestimation both gives binary weights to HMM-Sufficient Statistics prior to adaptation, where the N-best speakers have weight equal to one while the rest are zeros. Here, we will detail another technique meant to weight the HMM-Sufficient Statistics parameters prior to adaptation. In the speaker selection, each of the speaker in the N-best list has its corresponding likelihood score given the observation data. These likelihood scores are the basis of identifying the N-best among all of the speakers in the database. The intuition of weighting the HMM-Sufficient Statistics is to utilize these likelihood scores and based on these, introduce weighting mechanism of the individual Sufficient Statistics in the N-best list prior to adaptation. Figure 57 illustrates the idea of weighting. First, we have the test utterance as input to the system. N-best selection follows right after, where individual GMMs are used to get the individual likelihood given the test utterance. Next, we calculate the weights ω_p. These weights are then applied to the individual HMM-Sufficient Statistics as a mechanism to control its statistical components such as L_{im}^{spkr}, L_{ij}^{spkr}, $\boldsymbol{m}_{im}^{spkr}$, and $\boldsymbol{v}_{im}^{spkr}$. Weighting of the N-best HMM-Sufficient Statistics emphasizes the ones that are close to the test utterance while it attenuates those that are acoustically dissimilar. Lastly, we can use the newly weighted HMM-Sufficient

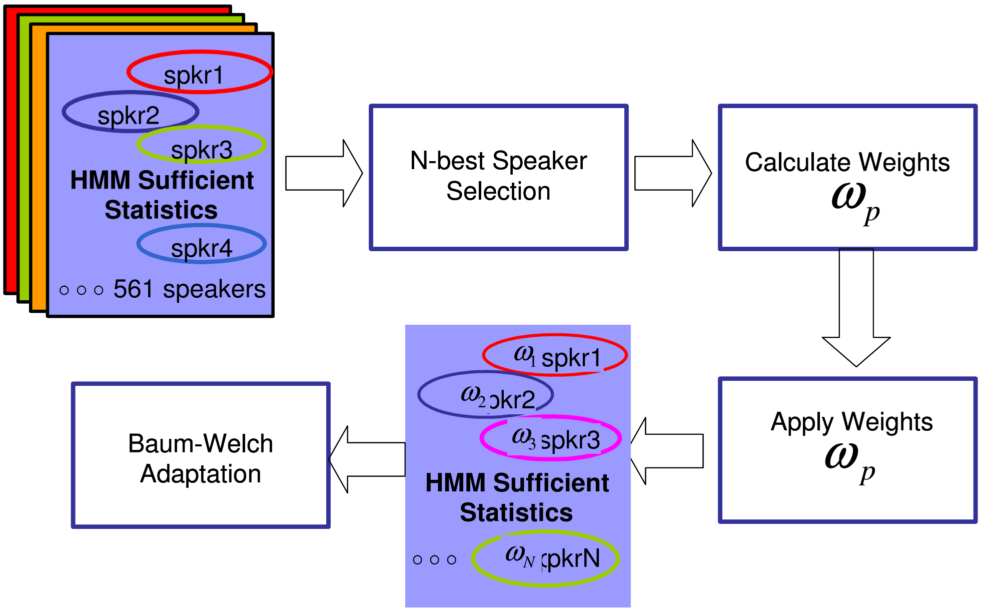


Figure 57. Block diagram of weighting HMM-Sufficient Statistics prior to adaptation.

Statistics as adaptation data.

We have provided two ways of calculating the weights, the first one is a linear weight which is independent of the likelihood scores. Its slope is dictated by the number or selected N-best. The linear weight is defined as

$$w_p = \frac{S-p}{S}, \tag{72}$$

where

$$\sum_{p=1}^{S} w_p = 1. \tag{73}$$

The second definition of the weight takes into consideration the likelihood scores from the N-best list and defined as

$$w_p = \frac{P(\mathrm{O}|\lambda_p)}{\sum_{s=1}^{S} P(\mathrm{O}|\lambda_s)} \tag{74}$$

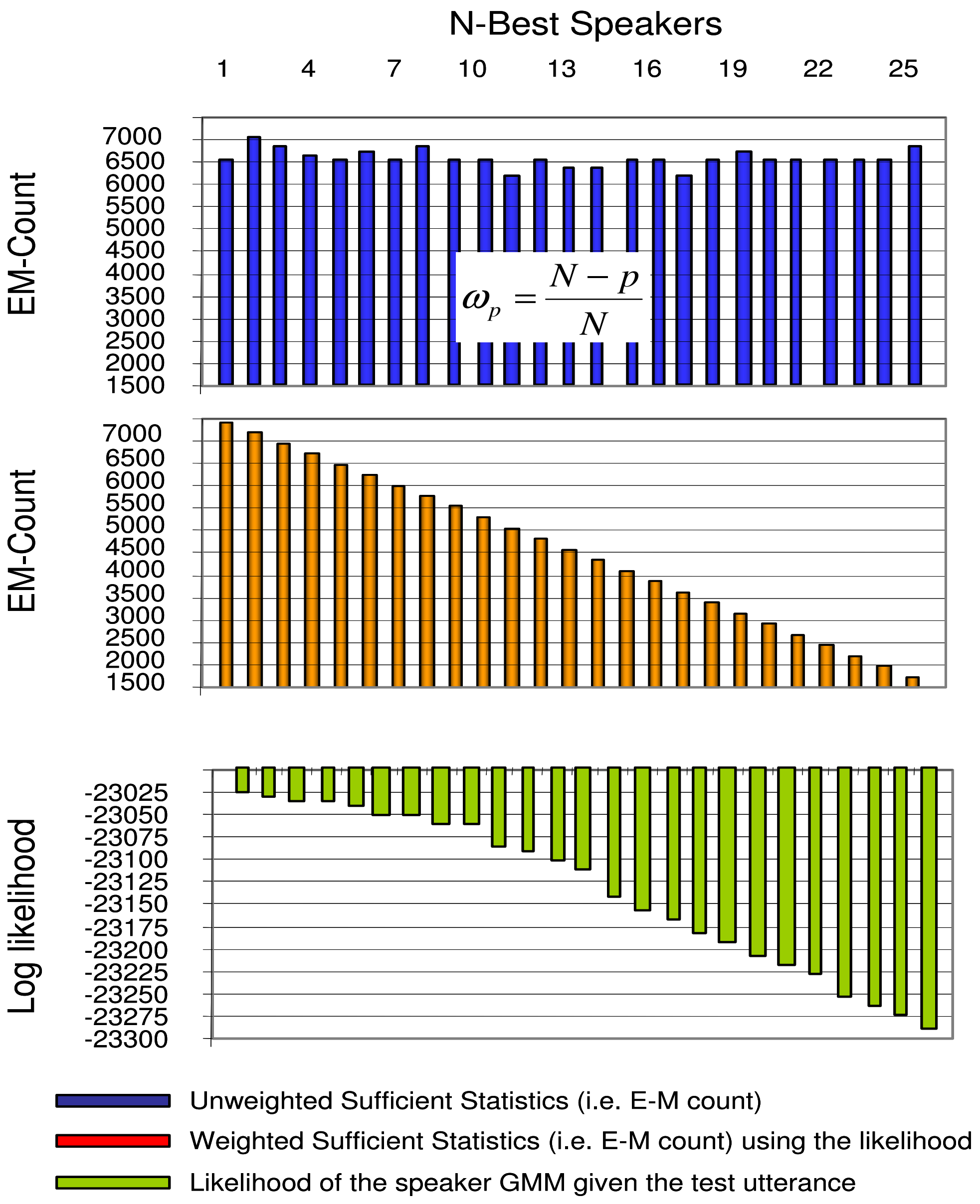


Figure 58. Effects of weighting the individual HMM-Sufficient Statistics using a linear weight.

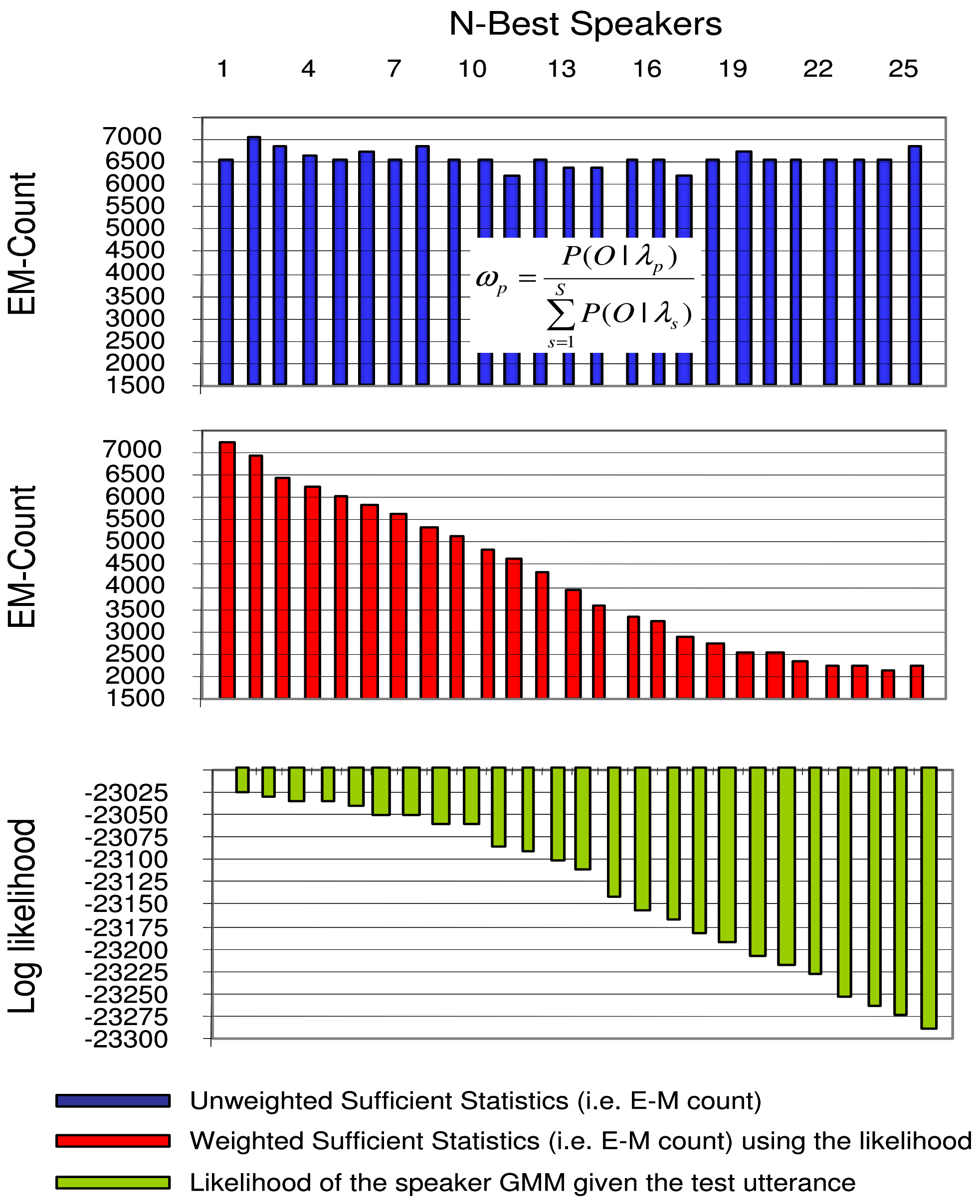


Figure 59. Effects of weighting the individual HMM-Sufficient Statistics derived from the likelihood of the GMMs given the test utterance.

where w_p is the corresponding weight of the *pth* speaker , $P(O|\lambda_p)$ is the likelihood of the observation O given the *pth* GMM model λ_p and S is the the number of selected speakers.

Figure 58 shows the mixture component occupancy of the selected N-best speakers (top) and its corresponding likelihood scores (bottom). On top, the light shaded bars are the unweighted mixture component occupancy (HMM-Sufficient Statistics) which is flat in general over N-best while the dark bars represent its weighted version using linear weighting. Furthermore, the weighted HMM-Sufficient Statistics has a decreasing trend over N-best speakers towards the acoustically dissimilar speakers which is also of the same trend as that of the likelihood scores of the individual speaker as shown in the bottom. Although this kind of weighting does not have any direct relationship with the actual likelihood scores, the envelope of the weighted HMM-Sufficient Statistics reflect the overall trend of the likelihood. In Figure 59 however, we used the likelihood weighting given in Equation 74. Thus, the envelope of the weighted Sufficient Statistics has a direct relationship with the envelope in the bottom of the figure unlike in Figure 58 where it just reflect the trend but no direct relevance with the envelope.

5.3.1 Results

We performed recognition experiments which involve weighting of the Sufficient Statistics, and the result is given in Figure 60. In this figure, we show the two types of weighting, based on likelihood and the linear weighting given in Equations 74 and 72 respectively. We compared the result to the baseline using the rapid Baum-Welch multi-template reestimation using binary weighting. There is a minimal improvement in both linear and likelihood weighting as opposed to the baseline.

5.3.2 Limitations of Weighting

Although recognition performance is slightly improved when using the linear and likelihood weights, this is not so significant as compared to the baseline approach where binary weighting is implemented. Weighting does not reduce the selected N-best speakers. Moreover, it requires additional computation load to execute Equations 74 and 72, thus adaptation time reduction cannot be achieved.

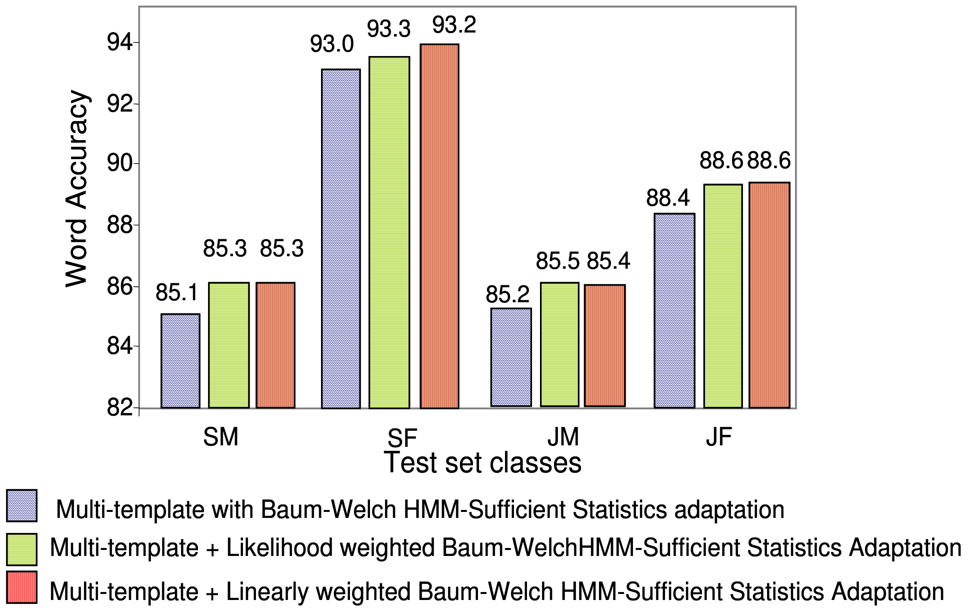


Figure 60. Recognition performance with the weighted HMM-Sufficient Statistics adaptation based on one-iteration of Baum-Welch.

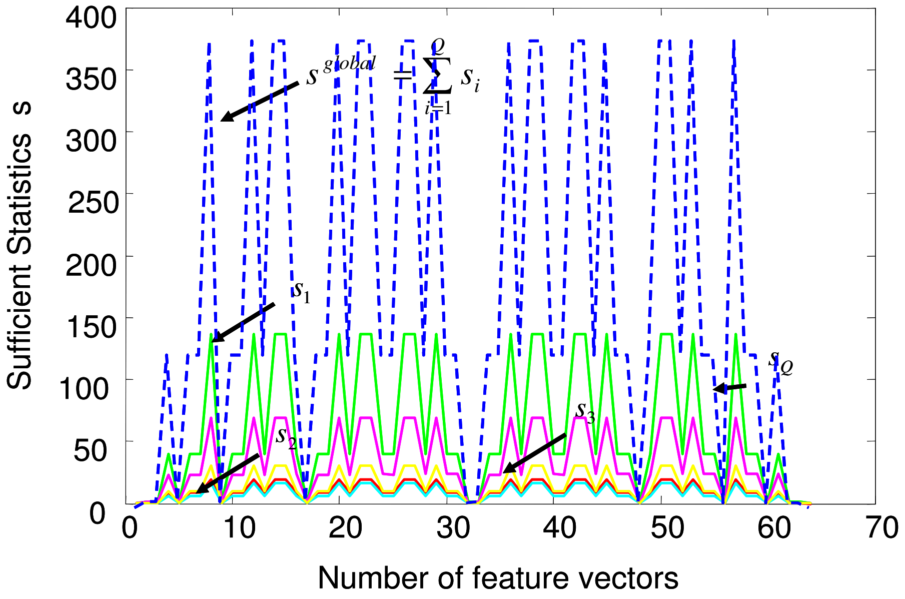


Figure 61. Illustration of the global Sufficient Statistics as the sum of all the speakers' Sufficient Statistics.

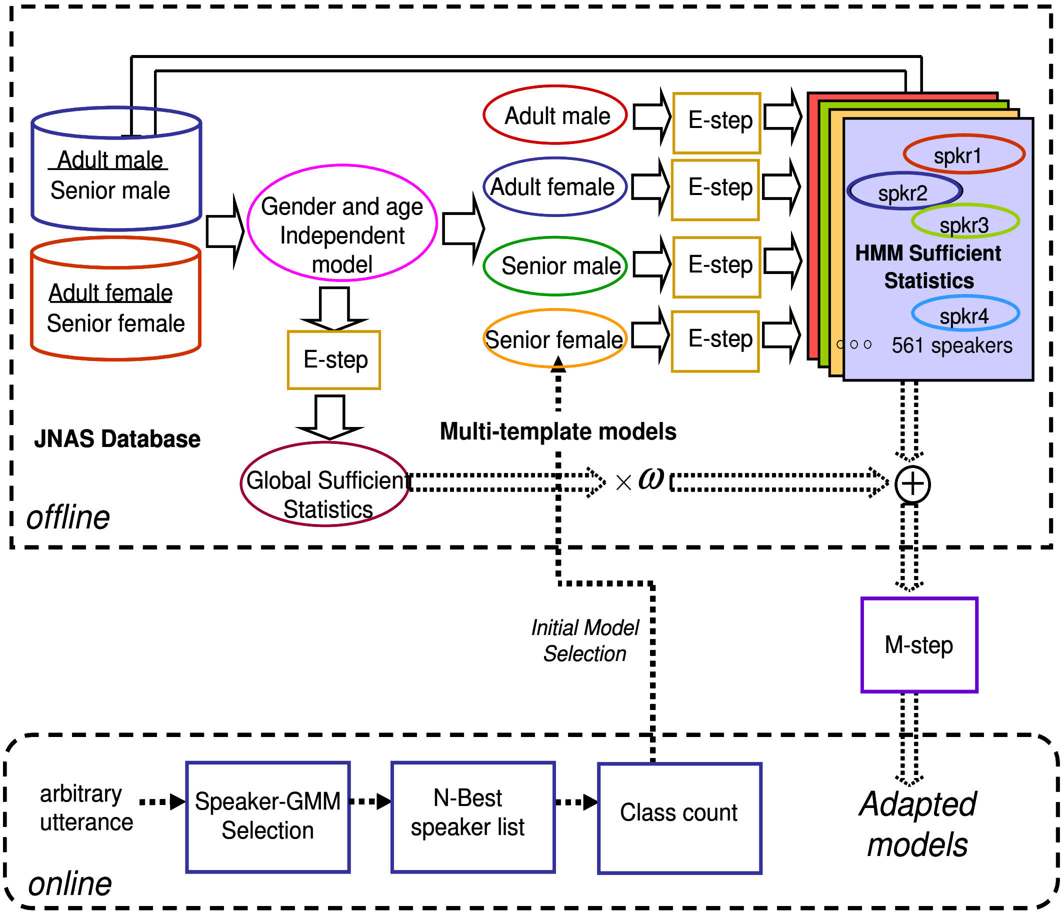


Figure 62. Block diagram of HMM-Sufficient Statistics adaptation with linear interpolation using individual speakers' Sufficient Statistics.

5.4 Linear Interpolation of the Global HMM-Sufficient Statistics and Clustering of Speakers

To further reduce adaptation time using the multi-template E-M approach, we introduce linear interpolation using the global Sufficient Statistics . In this section we discuss two types of implementing linear interpolation. First, treating all of the speakers in the database individually "individual speaker", the implementation that we have been using until now. The next implementation, is to impose clustering technique, replacing the individual speakers into clustered speakers.

To explain further the idea of global Sufficient Statistics, Figure 61 illustrates a global Sufficient Statistics denoted as s^{global} as a sum of all the speakers' Sufficient Statistics s^{spkr}. In particular, the global HMM-Sufficient Statistics are given as

$$\boldsymbol{m}_{im}^{global} = \sum\nolimits_{s=1}^{Q} \boldsymbol{m}_{im}^{s} \ , \tag{75}$$

$$\boldsymbol{v}_{im}^{global} = \sum\nolimits_{s=1}^{Q} \boldsymbol{v}_{im}^{s} \ , \tag{76}$$

$$L_{im}^{global} = \sum\nolimits_{s=1}^{Q} L_{im}^{s} \ , \tag{77}$$

and

$$L_{ij}^{global} = \sum\nolimits_{s=1}^{Q} L_{ij}^{s} \tag{78}$$

where Q is the total number of speakers in the database. $\boldsymbol{m}_{im}^{global}$, $\boldsymbol{v}_{im}^{global}$, L_{im}^{global}, and L_{ij}^{global} are the global HMM-Sufficient Statistics.
In Figure 62 we show the over-all block diagram of the proposed method using the the weighting of the global Sufficient Statistics together with the multi-template Baum-Welch adaptation. In the *offline* part, together with the creation of the multi-template HMM-Sufficient Statistics from multi-template models, we create a global HMM-Sufficient Statistics using the SI model. The actual adaptation *online* follows exactly the same as the previous HMM-based approach except for the linear interpolation. The provision of linear interpolation makes it possible to robustly estimate the target speaker's HMMs even with N-best reduced ($S < S_{optimal}$) since the weighted global Sufficient Statistics offsets the negative effect of the removed statistical information [46] to update a robust model. The adapted HMM parameters are as follows :

$$C_{im}^{adp(temp)} = \frac{\sum\nolimits_{s=1}^{S} L_{im}^{s(temp)} + \omega L_{im}^{global}}{\sum\nolimits_{m=1}^{M} (\sum\nolimits_{s=1}^{S} L_{im}^{s(temp)} + \omega L_{im}^{global})} \ , \tag{79}$$

$$\boldsymbol{\mu}_{im}^{adp(temp)} = \frac{\sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im}^{s(temp)} + \omega \boldsymbol{m}_{im}^{global}}{\sum\nolimits_{s=1}^{S} L_{im}^{s(temp)} + \omega L_{im}^{global}} \ , \tag{80}$$

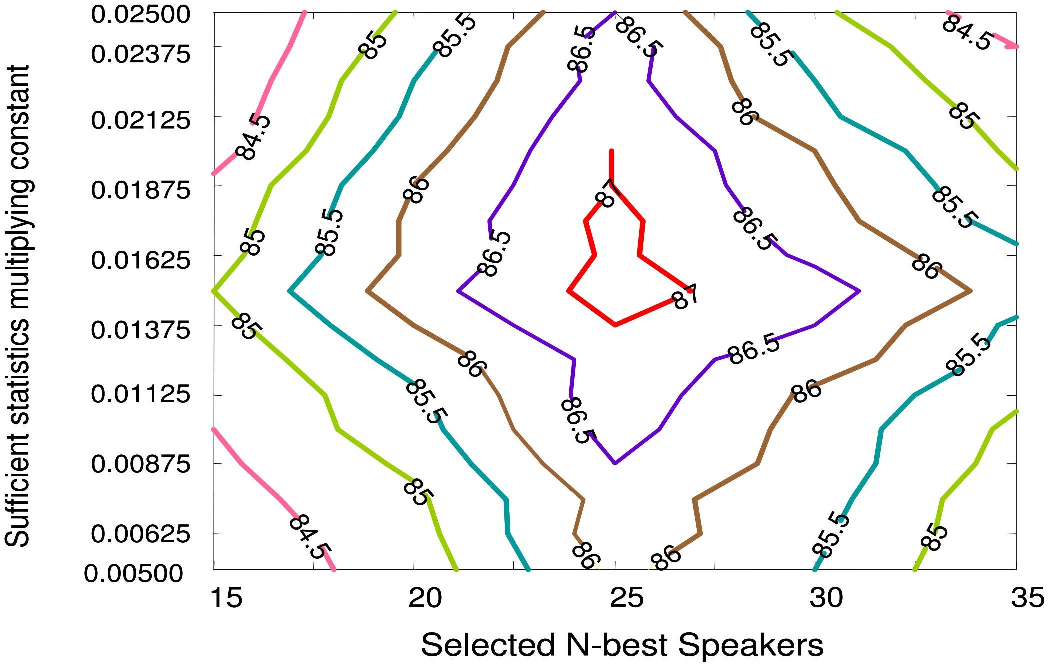


Figure 63. Contour plot of the weighting factor for interpolation.

$$\mathbf{\Sigma}_{im}^{adp(temp)} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s(temp)} + \omega \boldsymbol{v}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s(temp)} + \omega L_{im}^{global}} - \boldsymbol{\mu}_{im}^{adp(temp)} \boldsymbol{\mu}_{im}^{adp(temp)^T}, \tag{81}$$

and

$$a_{ij}^{adp(temp)} = \frac{\sum_{s=1}^{S} L_{ij}^{s(temp)} + \omega L_{ij}^{global}}{\sum_{j=1}^{J}(\sum_{s=1}^{S} L_{ij}^{s(temp)} + \omega L_{ij}^{global})} \tag{82}$$

where $C_{im}^{adp(temp)}$, $\boldsymbol{\mu}_{im}^{adp(temp)}$, $\mathbf{\Sigma}_{im}^{adp(temp)}$, and $a_{ij}^{adp(temp)}$ are the newly updated mixture weight, means, covariance matrix, and updated transition probability using linear interpolation. L_{im}^{s}, L_{ij}^{s}, $\boldsymbol{m}_{im}^{s}$, and $\boldsymbol{v}_{im}^{s}$ are the probability of mixture component occupancy, the accumulated probability of the state occupancy, means and variance respectively of the selected N-best speakers S. L_{im}^{global}, L_{ij}^{global}, $\boldsymbol{m}_{im}^{global}$, and $\boldsymbol{v}_{im}^{global}$ are the probability of the mixture occupancy, the accumulated probability of the state occupancy, means and variance respectively which are estimated using all of the training data which constitute the global Sufficient Statistics. ω

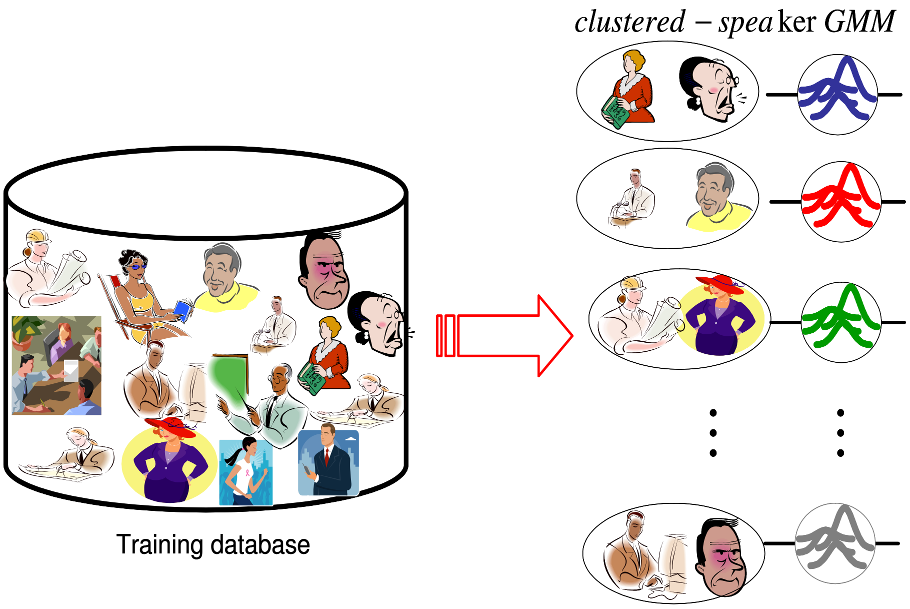


Figure 64. Creating clustered-speaker GMMs from the training database

is the weighting factor of the global HMM-Sufficient Statistics . Figure 63 shows the contour plot of the word accuracy (WA) at different values of the multiplying constant ω and corresponding N-best speakers selected. With the aid of this figure we set $\omega = 0.015$ with N-best=25.

In this part, we extended the linear interpolation approach by clustering the speakers in the database shown in Figure 65 as opposed to using only individual speakers in Figure 62. In this scheme, the individual-speaker GMMs are changed to cluster-based GMMs as shown in Figure **??**. Likewise, the individual HMM-Sufficient Statistics are changed to clustered speakers' HMM-Sufficient Statistics. The N-best list generates the list of clusters that are close to the target speaker. The motivation of this approach is to further reduce adaptation time by reducing N-best. Although, a further reduction of N-best poses a problem due to insufficient statistical data, this problem is minimized by combining 2 speakers' statistical information in each cluster and at the same time incorporate linear interpolation. In order to keep the statistical information uniform in the N-best list, we impose that each cluster be composed of a uniform number of speakers (ie. 2 speakers per cluster) by using Minimax [47].

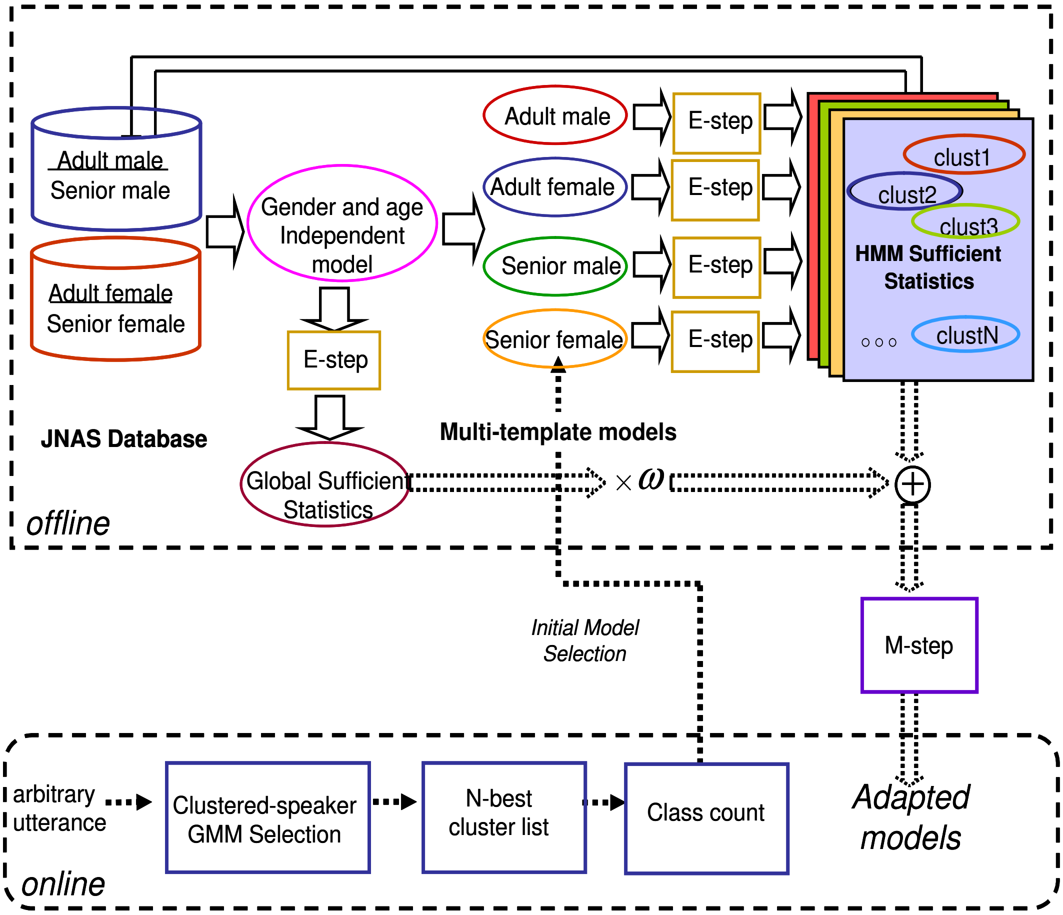


Figure 65. HMM-Sufficient Statistics adaptation with linear interpolation using clustered speakers' Sufficient Statistics.

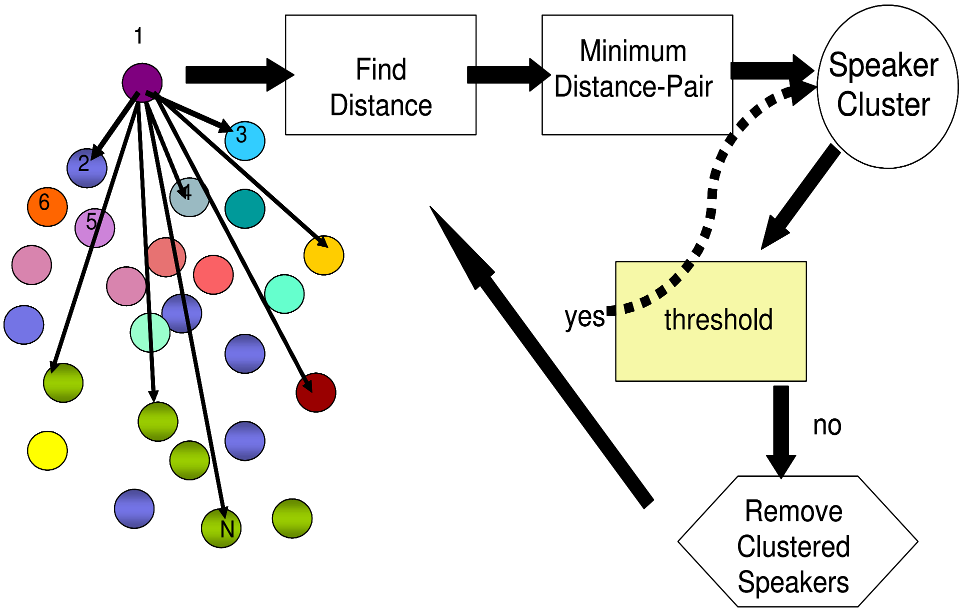


Figure 66. Clustering of speakers using Minimax.

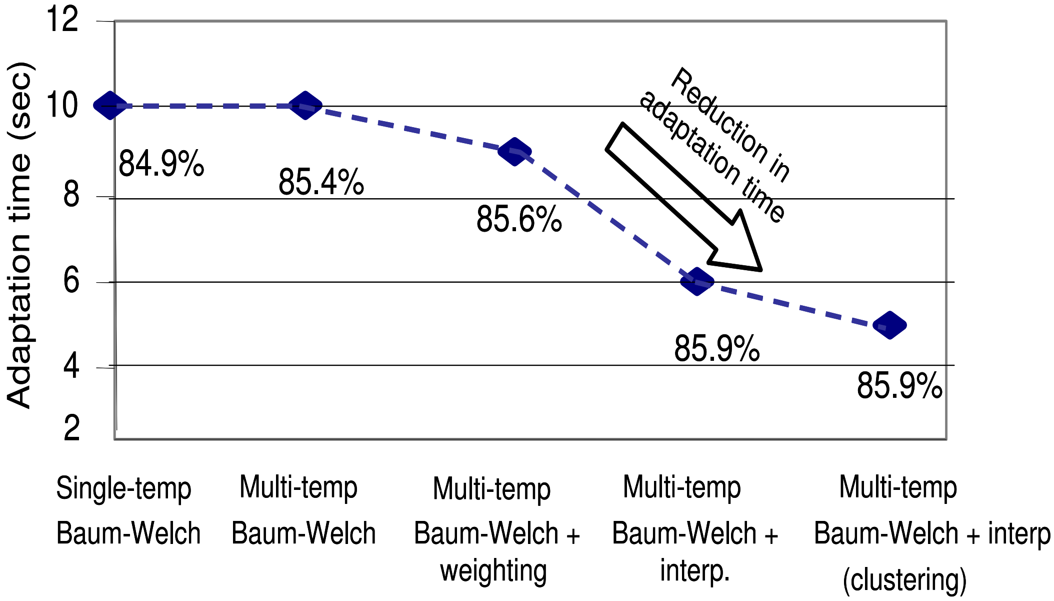


Figure 67. Reduction of adaptation time when using linear interpolation of the global HMM-Sufficient Statistics.

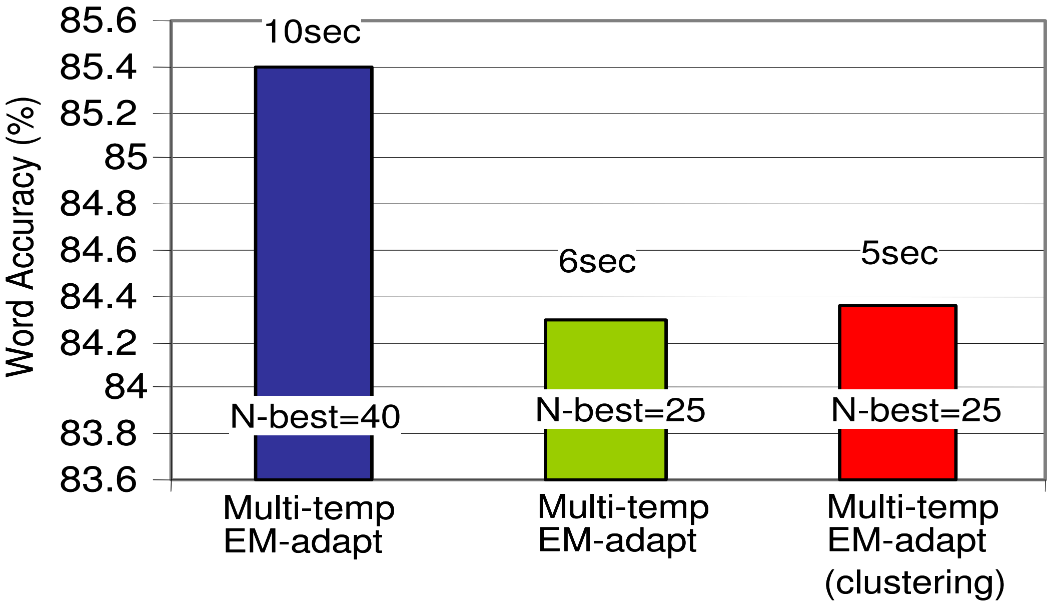


Figure 68. Recognition performance of a reduced N-best without interpolation of HMM-Sufficient Statistics.

Figure 66 illustrates the clustering procedure. First, the distances from a particular speaker relative to the rest of the speakers are calculated. This is done for each and every speaker in the speaker space and a table of distances is then generated for each speaker. From each of these table (each speaker), we find the minimum distance and create a new list which is composed of minimum distances for all speakers. From this new list we find the maximum distance. The speaker indices that has the maximum distance constitute a cluster and then removed in the speaker space. The process is iterated until all speakers are clustered. We choose to use this clustering technique in order to set two speakers per cluster unlike K-means clustering where the number of speakers per cluster are variable which has detrimental effects during model adaptation. Increasing the number of speakers per cluster results to a combination of speakers that are more acoustically dissimilar and this has a negative effect during adaptation. Experiment shows that by limiting each cluster to 2 speakers only, this negative effect is negligible. We also implemented K-means clustering but the former has a better recognition performance. The updated model parameters are given as

$$C_{im}^{adp(temp)} = \frac{\sum_{clust=1}^{S} L_{im}^{clust(temp)} + \omega L_{im}^{global}}{\sum_{m=1}^{M}(\sum_{clust=1}^{S} L_{im}^{clust(temp)} + \omega L_{im}^{global})}, \tag{83}$$

$$\boldsymbol{\mu}_{im}^{adp(temp)} = \frac{\sum_{clust=1}^{S} \boldsymbol{m}_{im}^{clust(temp)} + \omega \boldsymbol{m}_{im}^{global}}{\sum_{clust=1}^{S} L_{im}^{clust(temp)} + \omega L_{im}^{global}}, \tag{84}$$

$$\boldsymbol{\Sigma}_{im}^{adp(temp)} = \frac{\sum_{clust=1}^{S} \boldsymbol{v}_{im}^{clust(temp)} + \omega \boldsymbol{v}_{im}^{global}}{\sum_{clust=1}^{S} L_{im}^{clust(temp)} + \omega L_{im}^{global}} - \boldsymbol{\mu}_{im}^{adp(temp)} \boldsymbol{\mu}_{im}^{adp(temp)^T}, \tag{85}$$

and

$$a_{ij}^{adp(temp)} = \frac{\sum_{clust=1}^{S} L_{ij}^{clust(temp)} + \omega L_{ij}^{global}}{\sum_{j=1}^{J}(\sum_{clust=1}^{S} L_{ij}^{clust(temp)} + \omega L_{ij}^{global})} \tag{86}$$

where clust refers to the selected clustered speaker HMM-Sufficient Statistics.

5.4.1 Results

Figure 67 shows the result of interpolating the global HMM-Sufficient Statistics prior to one-iteration of Baum-Welch reestimation. Here we emphasize more, the time execution since this technique is designed to reduce adaptation time. In this graph, the adaptation time from the single-template Baum-Welch reestimation is significantly reduced from 10 sec to 5 sec as interpolation and clustering of speakers is used. Moreover there is no degradation of the recognition performance even though we reduced the adaptation data since it is compensated with the interpolation technique. However, Figure 68 shows a degradation in recognition performance when N-best is just merely reduced without using interpolation of the global HMM-Sufficient Statistics. Although adaptation time is reduced, the adapted model is not robust due to the insufficiency of adaptation data.

5.4.2 Limitations of Interpolation and Clustering

Although linear interpolation is effective in reducing the N-best speakers, its implementation in Baum-Welch makes it vulnerable to the limitations of the latter. Moreover, clustering of speakers may not be effective to a smaller number of speakers using a small database and combining speakers into clusters that are acoustically dissimilar will have negative effects during adaptation.

5.5 Summary

The HMM-Sufficient Statistics created offline is a result of the data reduction property of the Sufficient Statistics discussed in Chapter 3. This chapter discussed the utilization of the HMM-Sufficient Statistics as adaptation data. By using the N-best speakers' HMM-Sufficient Statistics, Baum-Welch reestimation is implemented online. Moreover we have introduced several variation of the rapid adaptation based on Baum-Welch from single-template to the multi-template approach. Weighting of the HMM-Sufficient Statistics is also discussed. We have also implemented linear interpolation of the global HMM-Sufficient Statistics together with the clustering of speakers where adaptation time was greatly reduced.

6. Rapid MLLR Unsupervised Speaker Adaptation Based on N-best Speakers' HMM-Sufficient Statistics

In the previous chapter we discussed the HMM-Sufficient Statistics adaptation based on Baum-Welch reestimation. Although this approach works very well, the system is still limited by the fact that Baum-Welch re-estimation is very sensitive to the amount of adaptation data being used. On the contrary, we aim at reducing adaptation data by reducing the number of N-best speakers. The more adaptation data being used, the more adaptation time is required. Reducing adaptation data implies a faster implementation. It is also important to note that we are not able to use the full potential of Baum-Welch re-estimation since we only allow a single iteration of the algorithm. Thus, it is imperative to use another adaptation scheme that is not sensitive to the size of adaptation data, can be implemented online, and delivers a better recognition performance than the one-iteration Baum-Welch reestimation. There is one adaptation scheme that fits the criteria, the MLLR adaptation approach is powerful and requires fewer adaptation data than Baum-Welch. However, this an offline approach and takes much more time than Baum-Welch owing to a more complex computational task if compared using the same amount of adaptation utterances. Moreover, it does not perform well when using only a single adaptation utterance.

In this chapter, we propose to extend the rapid Baum-Welch reestimation by using MLLR. By tailor-fitting this powerful algorithm to make use of N-best HMM-Sufficient Statistics adaptation, we can effectively reduce its current adaptation time comparable to that of the rapid Baum-Welch, to realize a rapid adaptation scheme. The importance of clustering acoustically close Gaussians in the form of regression tree which is important in the MLLR adaptation is also discussed. The concept of of MLLR mean and variance adaptation using the observation data will be presented together with its counterpart using HMM-Sufficient Statistics. This approach is similar to that of [2] [22], but extended to using the N-best speakers HMM-Sufficient Statistics as a mechanism to provide adaptation data and execute rapidly.

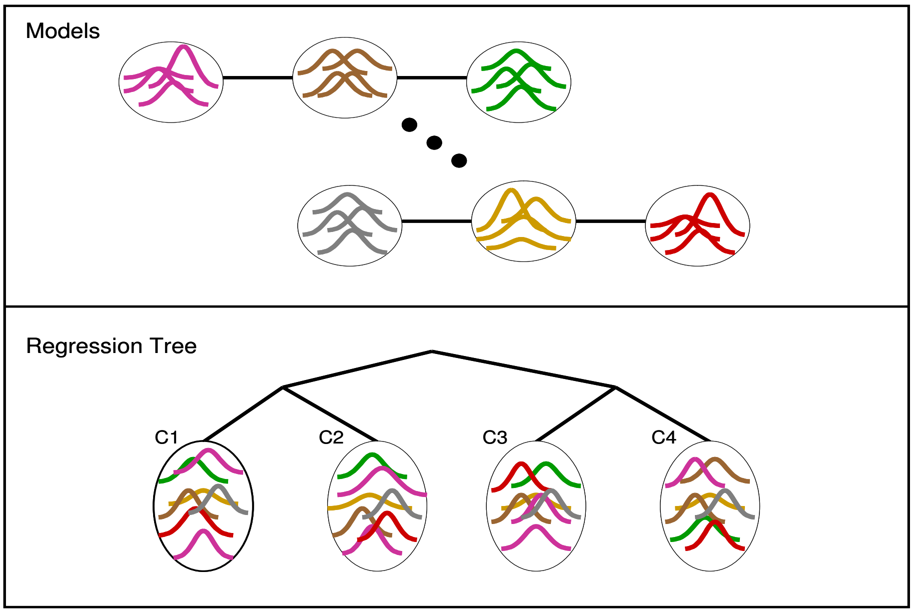


Figure 69. Mixture components of Gaussians from the model set that are acoustically close which are grouped together in the regression tree.

6.1 Regression Class Trees

Due to the enormous amount of models that we are dealing, it is improbable to get as much adaptation data that would effectively cover all of these using only several utterances. Thus, employing clustering of Gaussians through the regression tree would allow mixtures to be grouped and updated altogether. This is effective especially when adaptation data is scarce. Figure 69 illustrates the idea of creating regression tree for the model. In the uppermost part of the figure, the HMMs are shown with different Gaussian mixtures in every state. By imposing the regression tree as shown in the bottom of the figure, classes can be generated depending on the amount of the adaptation data. In the figure, four classes are generated and each of the class contains the corresponding mixtures that are acoustically close with each other. These mixtures may span across models and states. In effect, we keep the mixture-level of each of the speakers' HMM-Sufficient Statistics in the regression class with approximately 6.8MB in size. It is imperative to identify the class where Gaussians share the same transform.

As mentioned earlier, the choice of the number of classes depends primarily on the available adaptation data. The more adaptation data available, using more classes is recommended for a better discrimination of the groupings of Gaussian components. The regression class tree dictates the grouping of these Gaussians in the model set which allow for a possible adaptation of Gaussians even without adaptation data but proven to be acoustically close to some in which it is tied with.

6.2 Mean Adaptation

A global transformation is possible but with the size of N-best speakers adaptation data at hand we opt to find the individual transform at class level. In our approach, mean adaptation requires the calculation of the individual transformation matrix at the mixture level m in each class c. The adapted mean $\boldsymbol{\mu}_{m_c}^{adp}$ is given as

$$\boldsymbol{\mu}_{m_c}^{adp} = \boldsymbol{W}_{m_c}\boldsymbol{\zeta}_{m_c} \ , \tag{87}$$

where $\boldsymbol{W}_{m_c}$ is the unknown transformation matrix to be estimated that maximizes the likelihood of the adaptation data. The subscript m_c denotes that the transformation matrices are also tied across Gaussians. This allows a robust estimate of the transformations. $\boldsymbol{\zeta}_{m_c}$ is the extended mean vector of the HMM given as

$$\boldsymbol{\zeta}_{m_c} = [\omega \ \mu_1 \ \mu_2 \ \ldots \mu_n]^T \ . \tag{88}$$

In adapting the mean, the variance is unchanged thus we have this expression for the covariance matrix

$$\boldsymbol{\Sigma}_{m_c}^{adp} = \boldsymbol{\Sigma}_{m_c}, \tag{89}$$

we use E-M technique to solve for the maximization problem with the standard auxiliary function

$$\begin{aligned} Q(M, M^{adp}) = \frac{1}{2}\sum\nolimits_{r=1}^{R}\sum\nolimits_{m_c=1}^{M_c}\sum\nolimits_{t=1}^{T} L_{m_c}(t)[K^{(m)} + \log(|\boldsymbol{\Sigma}_{m_c}|) + \\ (\boldsymbol{o}(t) - \boldsymbol{\mu}_{m_c})^T \boldsymbol{\Sigma}_{m_c}^{adp-1}(\boldsymbol{o}(t) - \boldsymbol{\mu}_{m_c})]. \end{aligned} \tag{90}$$

where $K^{(m)}$ is a constant. After substituting Equation 87 to the auxiliary function and maximizing it we get

$$\boldsymbol{w}_{cp} = \boldsymbol{k}_c^{(p)} \boldsymbol{G}_c^{(p)-1} \ , \tag{91}$$

where $\boldsymbol{k}_c^{(p)}$ and $\boldsymbol{G}_c^{(p)-1}$ are given as

$$\boldsymbol{k}_c^{(p)} = \sum\nolimits_{m_c=1}^{M_c} \frac{1}{{\sigma_{m_c p}}^2} \boldsymbol{\zeta}_{m_c} {\boldsymbol{\zeta}_{m_c}}^T \sum\nolimits_{T=1}^{T} L_{m_c}(t) \tag{92}$$

and

$$\boldsymbol{G}_c^{(p)-1} = \sum\nolimits_{m_c=1}^{M_c} \sum\nolimits_{T=1}^{T} L_{m_c}(t) \frac{1}{{\sigma_{m_c p}}^2} \boldsymbol{o}_p(t) {\boldsymbol{\zeta}_{m_c}}^T . \tag{93}$$

In effect, mean adaptation evolves in the calculation of $\boldsymbol{k}_c^{(p)}$ and $\boldsymbol{G}_c^{(p)-1}$. It should be noted that in Equation 93, it is very important to have a direct access to the actual observation data $\boldsymbol{o}$.

6.3 Variance and Covariance Adaptation

The adapted variance and covariance are derived separately from the mean and is given by

$$\boldsymbol{\Sigma}_c^{adp} = \boldsymbol{B}_{m_c}^{\boldsymbol{T}} \boldsymbol{H}_c \boldsymbol{B}_{m_c} \ , \tag{94}$$

where we need to find linear transformation $\boldsymbol{H}$. $\boldsymbol{B}$ is the inverse of the Choleski factor $\boldsymbol{\Sigma}_{mr}^{-1}$ and expressed as

$$\boldsymbol{B} = \boldsymbol{C}_{m_c}^{-1} \ . \tag{95}$$

Substituting equation 94 to the auxiliary function and maximizing it, the linear transformation $\boldsymbol{H}_c$ is given as

$$\boldsymbol{H}_c = \frac{\sum\nolimits_{m_c=1}^{M_c} {\boldsymbol{C}_{m_c}}^T [L_{m_c}(t)(\boldsymbol{o}(t) - \boldsymbol{\mu}_{m_c}^{adp})(\boldsymbol{o}(t) - \boldsymbol{\mu}_{m_c}^{adp})^T] \boldsymbol{C}_{m_c}}{L_{m_c}(t)} . \tag{96}$$

Equation 96 is expressed in terms of the observation data $\boldsymbol{o}$, which signals its dependency to the actual collected adaptation utterances.

6.4 MLLR Adaptation Parameters Using HMM-Sufficient Statistics

In realizing a rapid MLLR adaptation, it is important to remove the dependency of the adaptation algorithm with the actual data samples $\boldsymbol{o}$. This is possible by virtue of HMM-Sufficient Statistics. This implies that the E-step is performed offline as discussed in Chapter 4. The new expressions for Equations 92, 93, and 96 are

$$\boldsymbol{k}_c^{(p)} = \sum\nolimits_{m_c=1}^{M_c} \frac{1}{\sigma_{m_c p}{}^2} \boldsymbol{\zeta}_{m_c} \sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} \ , \tag{97}$$

$$\boldsymbol{G}_c^{(p)-1} = \sum\nolimits_{m_c=1}^{M_c} \frac{1}{\sigma_{m_c p}{}^2} \boldsymbol{\zeta}_{m_c} \boldsymbol{\zeta}_{m_c}^{\boldsymbol{T}} \sum\nolimits_{s=1}^{S} L_{im_c}^{s} \ , \tag{98}$$

and

$$\boldsymbol{H}_c = \frac{\sum\nolimits_{m_c=1}^{M_c} \boldsymbol{C}_{m_c}^{\boldsymbol{T}} \boldsymbol{C}_{m_c}}{\sum\nolimits_{m_c=1}^{M_c} \sum\nolimits_{s=1}^{S} L_{im_c}^{s}} [\sum\nolimits_{s=1}^{S} \boldsymbol{v}_{im_c}^{s} \sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} \boldsymbol{\mu}_{m_c}^{adp\boldsymbol{T}} - \sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} \boldsymbol{\mu}_{m_c}^{adp} + \sum\nolimits_{s=1}^{S} L_{im_c}^{s} \boldsymbol{\mu}_{m_c}^{adp} \boldsymbol{\mu}_{m_c}^{adp\boldsymbol{T}}]. \tag{99}$$

These equations needed to carry out the MLLR adaptation are now functions of the selected N-best speakers' Sufficient Statistics $L_{im_c}^{s}$,$\boldsymbol{m}_{im_c}^{s}$, and $\boldsymbol{v}_{im_c}^{s}$ which are all precalculated parameters.

6.5 MLLR with Multi-template HMM-Sufficient Statistics

The block diagram of the proposed system is shown in Figure 70. This approach is still based on multi-template models in creating the HMM-Sufficient Statistics. During the actual adaptation process, we use MLLR adaptation instead of using the Baum-Welch approach in the M-step. This constitutes primarily the calculation of the parameters given by Equations 97-99. Since all adaptation data are already parameterized in advance, in the form of HMM-Sufficient Statistics, MLLR adaptation is executed in just a few seconds. In performing MLLR

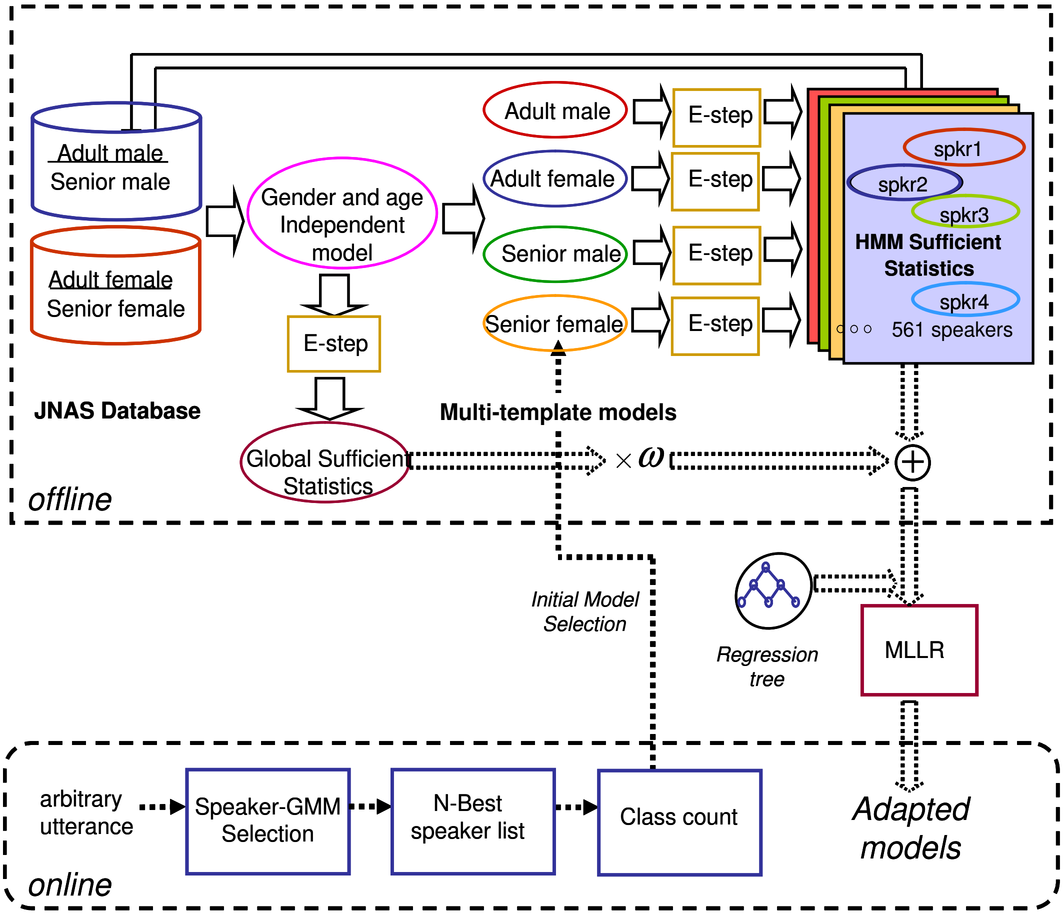


Figure 70. Block diagram of the online MLLR adaptation using HMM-Sufficient Statistics.

we obtain a set of transformation matrices that maximizes the likelihood of the adaptation data, given the model parameters.

6.5.1 Results

We evaluated the performance of the online MLLR approach using HMM-Sufficient Statistics. In Figure 71, the Baum-Welch based approach is shown together with the rapid MLLR based on HMM-Sufficient Statistics adaptation using 32, 64, and 128 classes. In this graph, the performance of the MLLR adaptation based on HMM-Sufficient Statistics improves as the number of classes is increased. This

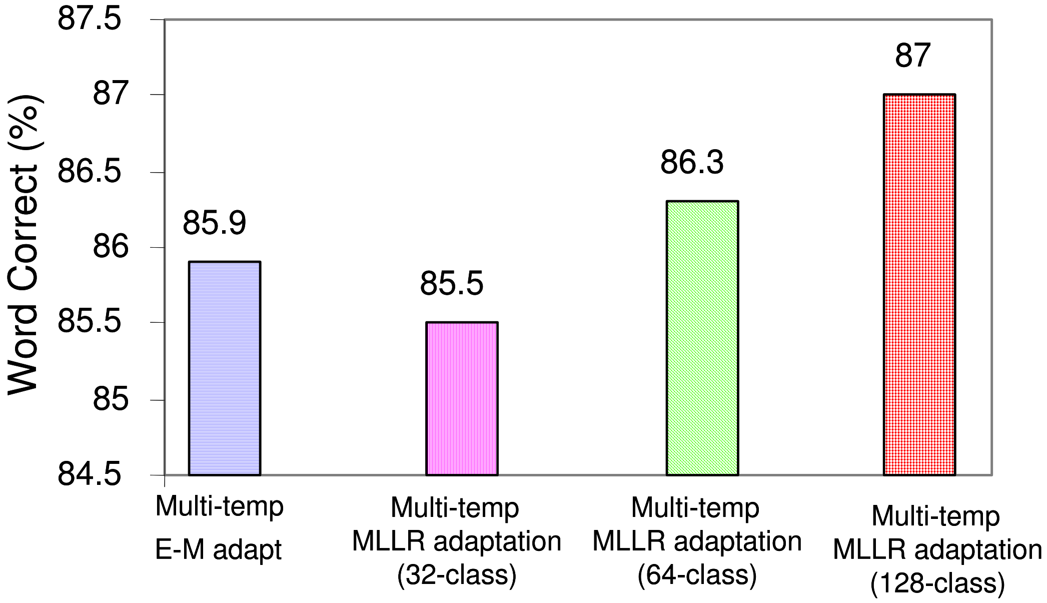


Figure 71. Recognition performance of the rapid MLLR adaptation based on HMM-Sufficient Statistics.

is logical since, we use many adaptation data. Lastly, it is important to note that the MLLR adaptation based on HMM-Sufficient Statistics outperforms the one-iteration Baum-Welch discussed in Chapter 4.

6.6 Linear Interpolation of the Global HMM-Sufficient Statistics in MLLR

We have extended the MLLR based HMM-Sufficient Statistics adaptation to using global interpolation of the HMM Sufficient Statistics. Linear interpolation of the global HMM Sufficient Statistics result to a reduction in adaptation time in the Baum-Welch approach in Chapter 5. Here, we will apply interpolation in the online MLLR adaptation. With this, it may be possible to reduce the N-best used in the online MLLR without degrading recognition performance as this will be compensated by the interpolation of the global HMM Sufficient Statistics. The parameters needed for mean adaptation are

$$\boldsymbol{k}_c^{(p)} = \sum\nolimits_{m_c=1}^{M_c} \frac{1}{{\sigma_{m_c p}}^2} \boldsymbol{\zeta}_{m_c} \sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} + \omega \boldsymbol{m}_{im_c}^{global} \tag{100}$$

and

$$\boldsymbol{G}_c^{(p)-1} = \sum\nolimits_{m_c=1}^{M_c} \frac{1}{{\sigma_{m_c p}}^2} \boldsymbol{\zeta}_{m_c} \boldsymbol{\zeta}_{m_c}^{\boldsymbol{T}} \sum\nolimits_{s=1}^{S} L_{im_c}^{s} + \omega L_{im_c}^{global} \ , \tag{101}$$

where ω is the interpolating factor of the global HMM Sufficient Statistics. Moreover, the parameter needed for the variance adaptation is

$$\begin{aligned}\boldsymbol{H} = {} & \frac{\sum\nolimits_{m_c=1}^{M_c} \boldsymbol{C}_{m_c}^{\boldsymbol{T}} \boldsymbol{C}_{m_c}}{\sum\nolimits_{m_c=1}^{M_c} \sum\nolimits_{s=1}^{S} L_{im_c}^{s}} [(\sum\nolimits_{s=1}^{S} \boldsymbol{v}_{im_c}^{s} + \omega \boldsymbol{v}_{im_c}^{global})(\sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} + \omega \boldsymbol{m}_{im_c}^{global}) \boldsymbol{\mu}_{m_c}^{adp\boldsymbol{T}} \\ & - (\sum\nolimits_{s=1}^{S} \boldsymbol{m}_{im_c}^{s} + \omega \boldsymbol{m}_{im_c}^{global}) \boldsymbol{\mu}_{m_c}^{adp} + (\sum\nolimits_{s=1}^{S} L_{im_c}^{s} + \omega L_{im_c}^{global}) \boldsymbol{\mu}_{m_c}^{adp} \boldsymbol{\mu}_{m_c}^{adp\boldsymbol{T}}].\end{aligned} \tag{102}$$

6.6.1 Results

We have successfully reduced adaptation time when using linear interpolation in the Baum-Welch implementation, and in Figure 72 we show the result when applying interpolation in the online MLLR adaptation. In this figure we show the adaptation time for all classes. The dashed line represents the normal MLLR online adaptation without interpolation while the block line shows the result when interpolation is employed. In the case of the 32-class, interpolation does not reduce adaptation time due to poor discrimination of classes. When classes are increased to 64 and 128 owing to a considerable amount of adaptation data available, a reduction of one second is possible with the 64-class and 128-class is achieved without degradation of the recognition performance. This result shows that the linear interpolation technique implemented to the rapid Baum-Welch reestimation is robust, and can be applied to MLLR as well.

6.7 Summary

We have just successfully implemented another HMM-Sufficient Statistics adaptation using MLLR [4] which is a far more effective adaptation scheme than the

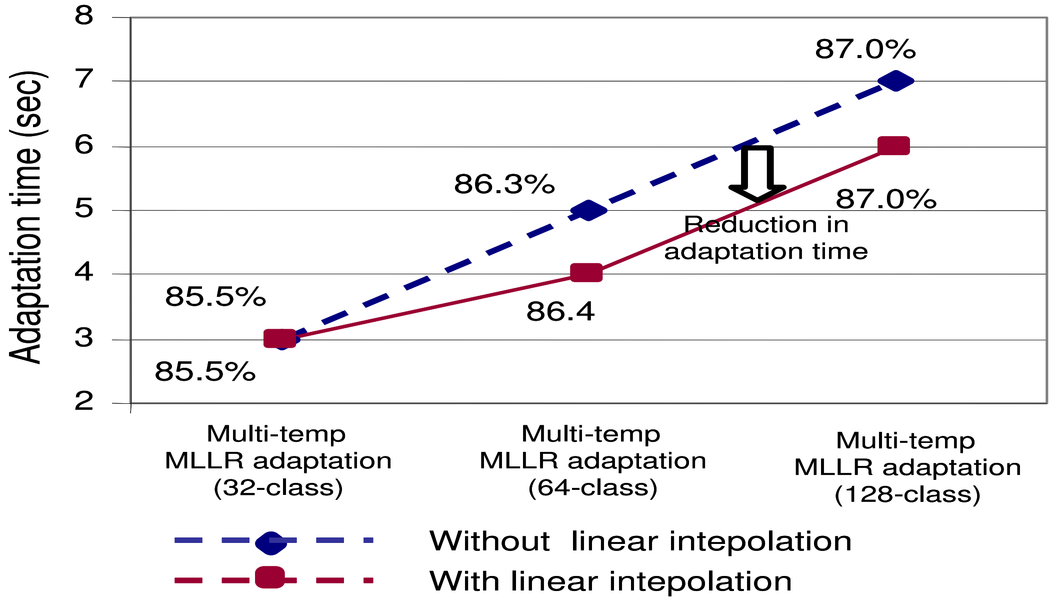


Figure 72. Reduction of adaptation time when using linear interpolation of the global HMM-Sufficient Statistics in the online MLLR adaptation.

Baum-Welch reestimation. By redesigning the conventional MLLR to accommodate the N-best speakers' HMM-Sufficient Statistics as adaptation data, a rapid implementation of MLLR is achieved using only one arbitrary utterance. The rapid MLLR adaptation which is the ultimate adaptation technique of this thesis has better recognition performance than that of the Baum-Welch approach. We also show that reduction of adaptation time is possible when the linear interpolation is combined with the online MLLR HMM-Sufficient Statistics adaptation.

7. Experimental Results and Discussion

In this chapter, we will discuss the basic setup, conditions and parameters used in the experiments. To show the robustness of the rapid adaptation using HMM-Sufficient Statistics for both Baum-Welch and MLLR, different types of noise are considered like office, car, crowd, and booth noise. Tests are also conducted in different SNRs from 10 dB, 15 dB, 20 dB, and 25 dB. Moreover, a more comprehensive discussion of the advantages using the rapid approach as compared to the existing adaptation schemes like MAP, MLLR and VTLN.

7.1 Experimental Setup

In the acoustic modelling part, 25 dB office noise is superimposed to the speech database [44] in creating the phonetically tied mixture (PTM) HMMs [48]. In the adaptation part, the single arbitrary noisy utterance is denoised with SS which is used for speaker selection. Lastly, for the actual recognition testing, the SS-denoised test utterances are superimposed with 30 dB office noise prior to recognition to neutralize the residual noise [44]. Figure 73 shows the overall block diagram of the system.

The test set is composed of four classes, namely: Adult male, Adult female, Senior male and Senior female. Each class is of 100 utterances from 23 speakers which are taken outside of the training speakers. This sums up to 400 total test utterances from 92 test speakers across different genders and age-groups. The speakers used in testing are different speakers from that of the training database. Recognition experiments are carried out using JULIUS [1] with 20K-word on Japanese newspaper dictation task from JNAS. The language model is provided by the IPA dictation toolkit. Table 1 summarizes the recognition parameters used in our experiment.

All of the speakers in the Adult and Senior JNAS training databases have four HMM-Sufficient Statistics which account the four multi-template HMMs (Adult Male, Adult Female, Senior Male and Senior Female) except for the single-template approach. The databases is consist of 60K-utterance from 301 (JNAS Adult) speakers and 53K-utterance from 260 (JNAS Senior) speakers where each speaker has 150 and 200 utterances respectively [6] as shown in Table 2. The size

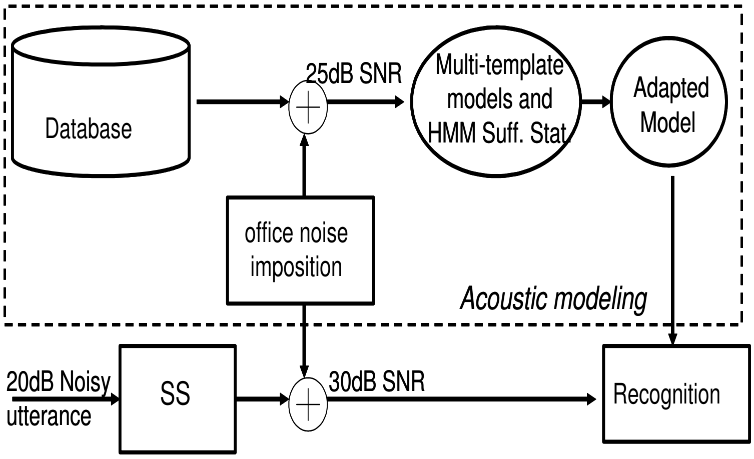


Figure 73. Block diagram of the overall system implementation.

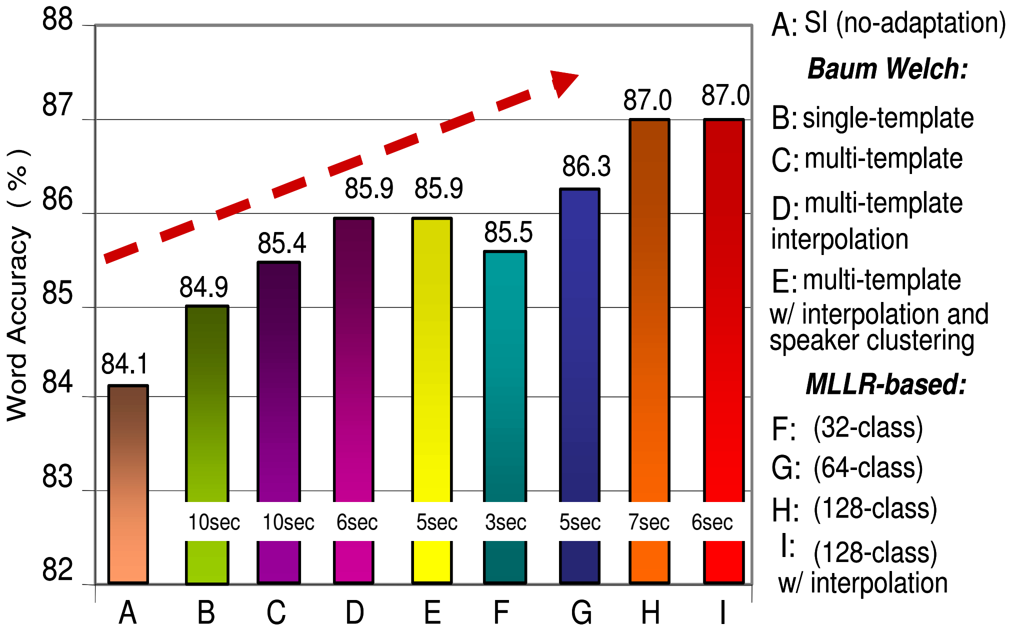


Figure 74. Overall recognition performance for 25 dB office noise environment using all platforms of HMM-Sufficient Statistics adaptation.

Table 1. System specifications

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior by JNAS
Test data	Adult and Senior by JNAS

Table 2. Set-up of Adult and Senior JNAS Database for HMM-Sufficient Statistics

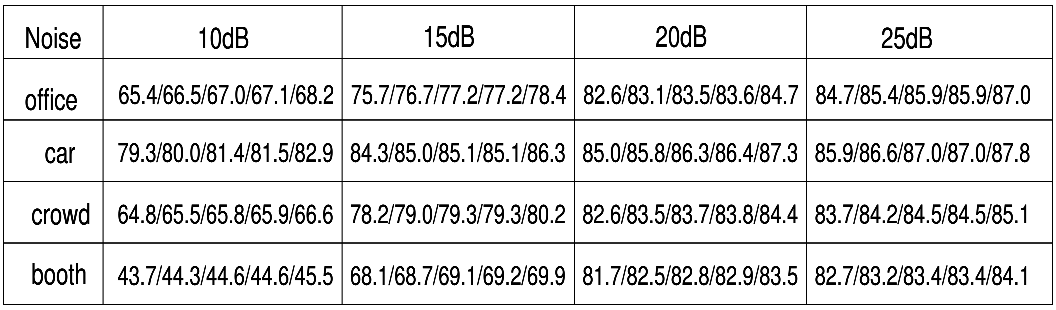
Database	Gender	Speakers	Utterances
Adult JNAS	Male	151	150
Adult JNAS	Female	150	150
Senior JNAS	Male	130	200
Senior JNAS	Female	130	200

Table 3. HMM-Sufficient Statistics Info

Rapid adaptation	HMM-Suff Stat per speaker	Size per Suff Stat
Baum-Welch: single-temp	1	5.5MB
Baum-Welch: multi-temp	4	5.5MB
Baum-Welch: Interpolation	4	5.5MB
MLLR: multi-temp	4	6.8MB
MLLR: Interpolation	4	6.8MB

of the individual HMM-Sufficient Statistics is approximately 5.5MB for the Baum-Welch and 6.8MB for the rapid MLLR which are stored in the disk. Summary of the HMM-Sufficient Statistics is given in Table 3.

Table 4. Word Accuracy of various SNR Statistics (Single-temp: $S_{optimal} = 40$ / Multi-temp: $S = 40$ / Multi-temp with interpolation $S = 25$ /Multi-temp with interpolation, Clustered-Speakers $S = 20$ / MLLR-based 128-class.



Noise	10dB	15dB	20dB	25dB
office	65.4/66.5/67.0/67.1/68.2	75.7/76.7/77.2/77.2/78.4	82.6/83.1/83.5/83.6/84.7	84.7/85.4/85.9/85.9/87.0
car	79.3/80.0/81.4/81.5/82.9	84.3/85.0/85.1/85.1/86.3	85.0/85.8/86.3/86.4/87.3	85.9/86.6/87.0/87.0/87.8
crowd	64.8/65.5/65.8/65.9/66.6	78.2/79.0/79.3/79.3/80.2	82.6/83.5/83.7/83.8/84.4	83.7/84.2/84.5/84.5/85.1
booth	43.7/44.3/44.6/44.6/45.5	68.1/68.7/69.1/69.2/69.9	81.7/82.5/82.8/82.9/83.5	82.7/83.2/83.4/83.4/84.1

7.2 Basic Results using HMM-Sufficient Statistics Adaptation

Recognition performance and the corresponding adaptation time of the rapid HMM-Sufficient Statistics adaptation both using Baum-Welch and MLLR is shown in Figure 74. In (A), the WA when using SI (no-adaptation) is 84.1%, which is improved to 84.9% when Baum-Welch based single template HMM-Sufficient Statistics adaptation is used with 10 sec adaptation time. Consequently, WA increased to 85.4% when multi-template approach is implemented (C). Since we only used a single iteration of Baum-Welch, the system's ability to select an acoustically close model through the multi-template implementation, improved the system's performance. Its corresponding adaptation time is left unchanged to 10 sec. Linear Interpolation of the global HMM-Sufficient Statistics in (D) significantly reduced the adaptation time to 6 sec. This approach compensates the effect of a degradation of the recognition performance when using fewer adaptation data by decreasing N-best from $S = 40$ to $S = 25$. In fact, the recognition performance slightly increased as compared to (C). When speakers are clustered prior to the creation of the HMM-Sufficient Statistics together with linear interpolation (E), the recognition performance remained unchanged in (D) but a further decrease of the adaptation time from 6 sec to 5 sec is achieved.

In the case of using MLLR for the rapid HMM-Sufficient Statistics adaptation as discussed in Chapter 5, a recognition performance of 85.5% is attained when using only 32 classes. There is a slight decrease in WA compared to the Baum-Welch reestimation with linear interpolation in (E). However, adaptation time is reduced to 3 sec from 5 sec. As we increased the number of classes to 64 (G), WA increased to 86.3% which is better than that of (E). When we used 128 classes, we have achieved 87.0% WA with 7 sec adaptation time. This has an absolute improvement of 1.1% compared to the best-performing Baum-Welch based approach (D) and (E). Lastly, when applying the interpolation approach for the online MLLR using 129 classes (I), we have maintained the recognition performance at 87.0% the same as in (H) but adaptation time is reduced to 6 seconds.

In Table 4, the summary of recognition performance in office, crowd, car and booth noise environments with different SNRs are given. In this result here, we use the optimal N-best result that corresponds to the highest recognition performance of a particular set-up. By looking at the WA we can see the improvement of the recognition performance of the various HMM-Sufficient Statistics schemes from the conventional single-template of the Baum-Welch based, up to the online MLLR approach.

7.3 Adaptation Time of Conventional MLLR

The conventional MLLR that makes use of the actual observed adaptation data consumes more time than the rapid MLLR that takes as input the N-best speakers' HMM-Sufficient Statistics. The fact that gathering and transcribing these utterances play an integral part in the adaptation process. In Figure 75, a graph of adaptation time as a function of N-best is shown when using the conventional MLLR approach taking as input the actual observed data. We compare this adaptation time with the rapid MLLR using HMM-Sufficient Statistics as adaptation data. The horizontal broken line shows the amount of adaptation needed for the conventional approach and the rapid approach. With N-best=11, adaptation time are as follows : 100 sec (32-class), 130 sec (64-class) and 200 sec (128-class). These shows that adaptation time is significantly higher compared to the proposed online MLLR adaptation based on HMM-Sufficient Statistics using

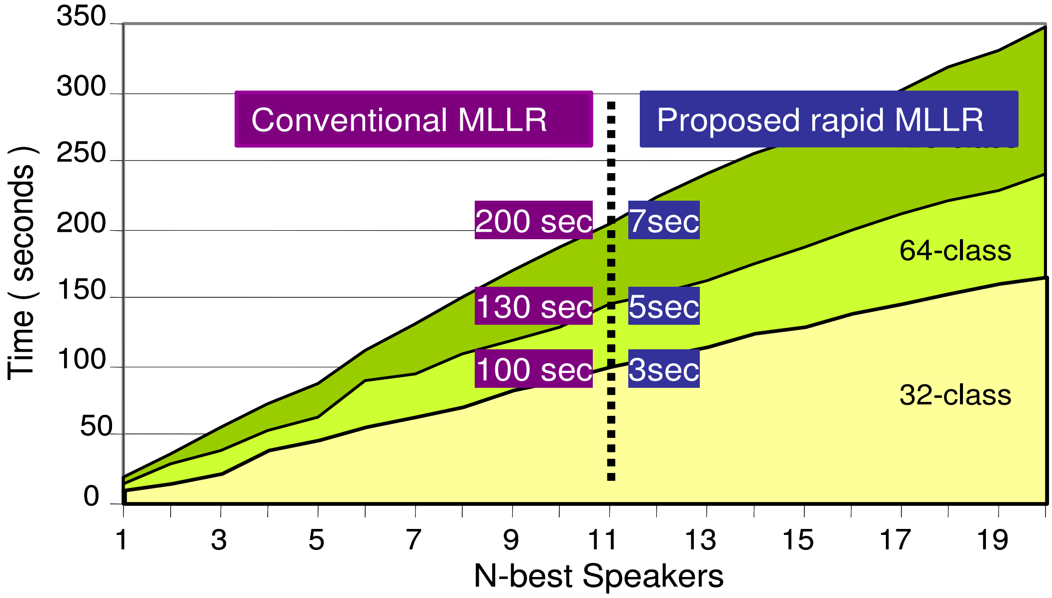


Figure 75. Adaptation time of the conventional MLLR and the proposed MLLR adaptation based on HMM-Sufficient Statistics.

the same N-best=11 as shown in Figure 74 with the following adaptation time: 3 sec (32-class), 5 sec (64-class) and 7 sec (128-class). The very low adaptation time for the HMM-Sufficient Statistics based MLLR adaptation is attributed to the fact that statistical parameters are already pre-calculated and are readily available for adaptation.

7.4 Detailed Results on Speaker Clustering and Linear Interpolation

In Chapter 4, the clustering of speakers combined with the interpolation of the global HMM-Sufficient Statistics is the bets performing rapid Baum-Welch adaptation. In this scheme, the adaptation time has been reduced without degrading the recognition performance even though adaptation data is reduced. In this section, a more detailed result on clustering of speakers together with linear interpolation is presented. Figure 76 shows the plot of the WA comparing in detail when using 1) individual speakers (unclustered) with interpolation, 2) clustered

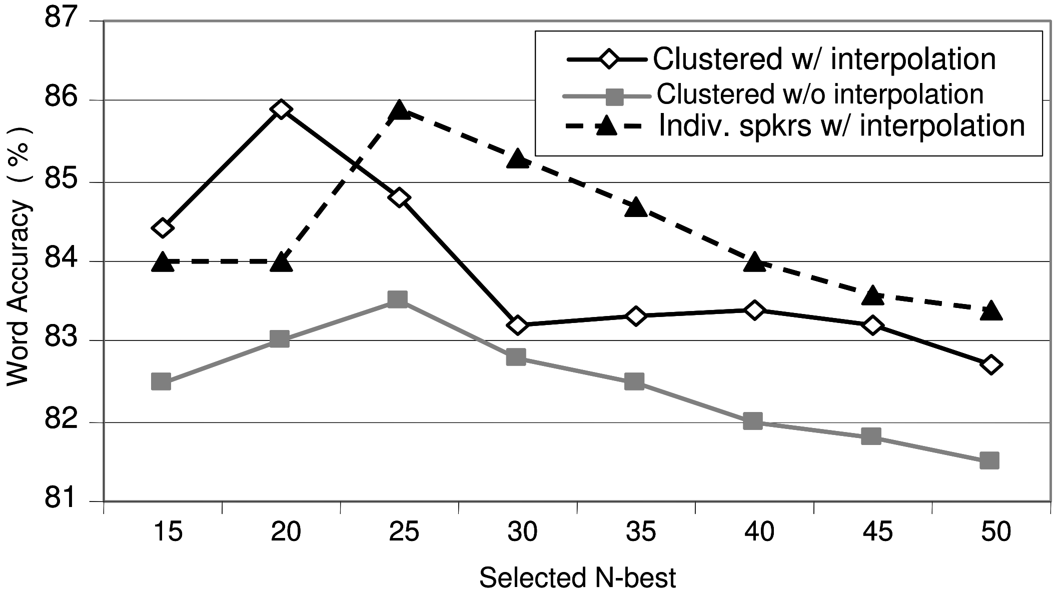


Figure 76. Recognition Performance of the Baum-Welch based linear interpolation approach.

speakers with and without linear interpolation as a function of N-best. The N-best list for the unclustered speakers are the original speakers in the database while the latter's' N-best list is composed of two speakers, clustered together. In this graph, it is apparent that the proposed linear interpolation improves the performance of the clustered speakers as opposed to the clustered speakers Baum-Welch adaptation without linear interpolation. More interestingly, the clustered speakers with linear interpolation using N-best =20 achieved the same recognition performance with that of using the individual speakers (unclustered) with N-best = 25, thus a reduction in adaptation time is further attained. The effect of the clustering of speakers is manifested in the reduction of adaptation time since N-best is further reduced. In the case of clustering, it is possible that the two speakers are dissimilar and when clustered altogether this will have a negative impact in the adaptation process. The linear interpolation is responsible of compensating for the reduction adaptation data as a consequence of dropping some clusters in the N-best list.

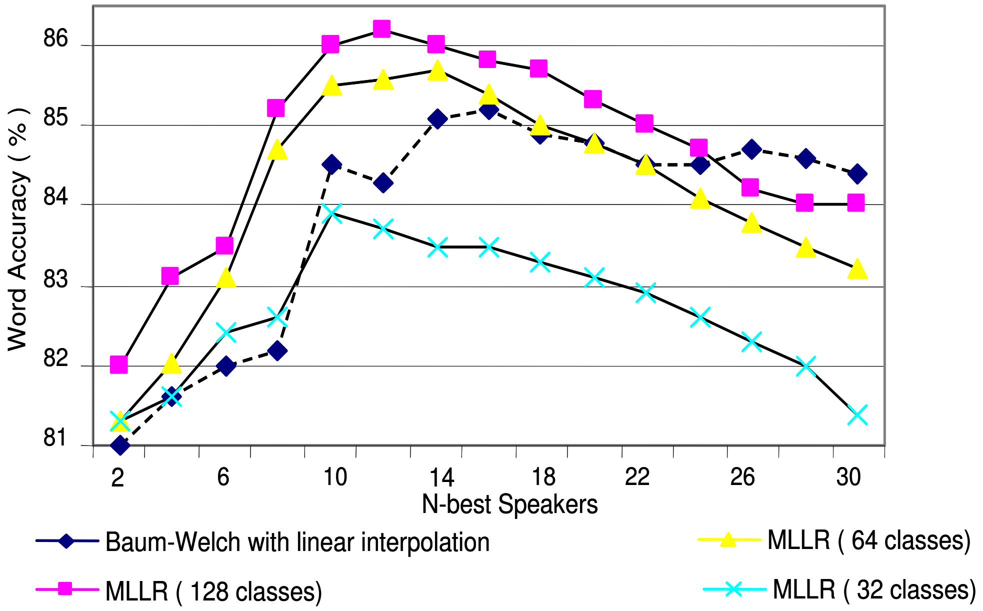


Figure 77. Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Adult Male).

7.5 Detailed Comparison Between the Rapid Baum-Welch and Rapid MLLR based Adaptation

In this section, the Baum-Welch adaptation with linear interpolation, which is the best-performing HMM-Sufficient Statistics adaptation in Chapter 5 is compared with the rapid MLLR approach as shown in Figures 77, 78, 79, and 80. These figures we show the recognition performance in every class (Adult male, Senior male, Adult female, and Senior female). We can observe that when using the Baum-Welch based approach, the system needs more N-best to reach the optimum recognition performance while it takes fewer N-best for the online MLLR approach. This result points to the fact that online MLLR is very robust to smaller adaptation data than Baum-Welch. Moreover, it is also clear that once the online MLLR approach reaches the optimal recognition performance, further increasing N-best from that point onwards manifests a decreasing trend of the recognition performance. This is attributed to the fact that we are using different speakers' HMM-Sufficient Statistics . Whenever N-best is increased, acoustical

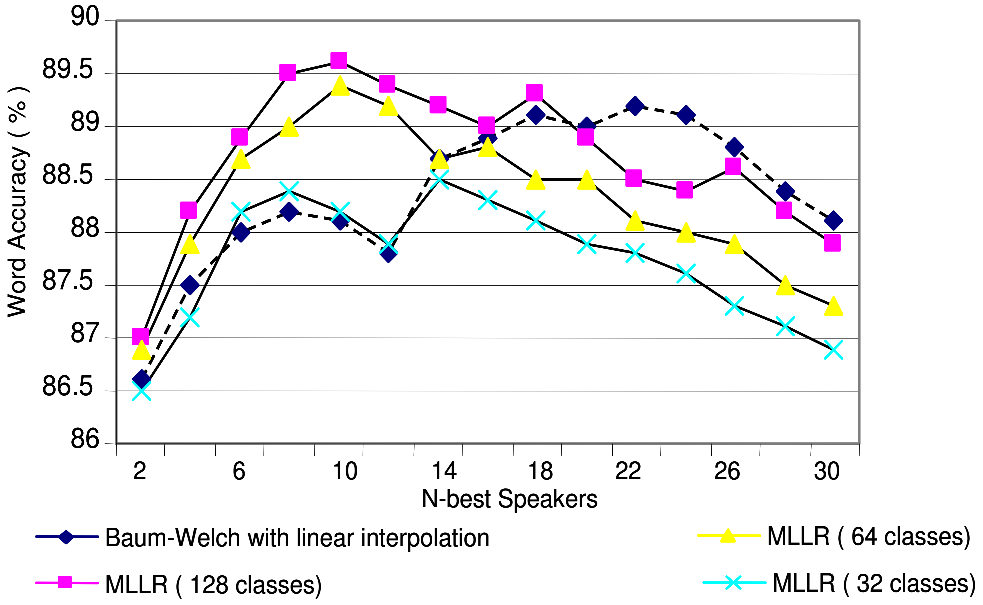


Figure 78. Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Adult Female).

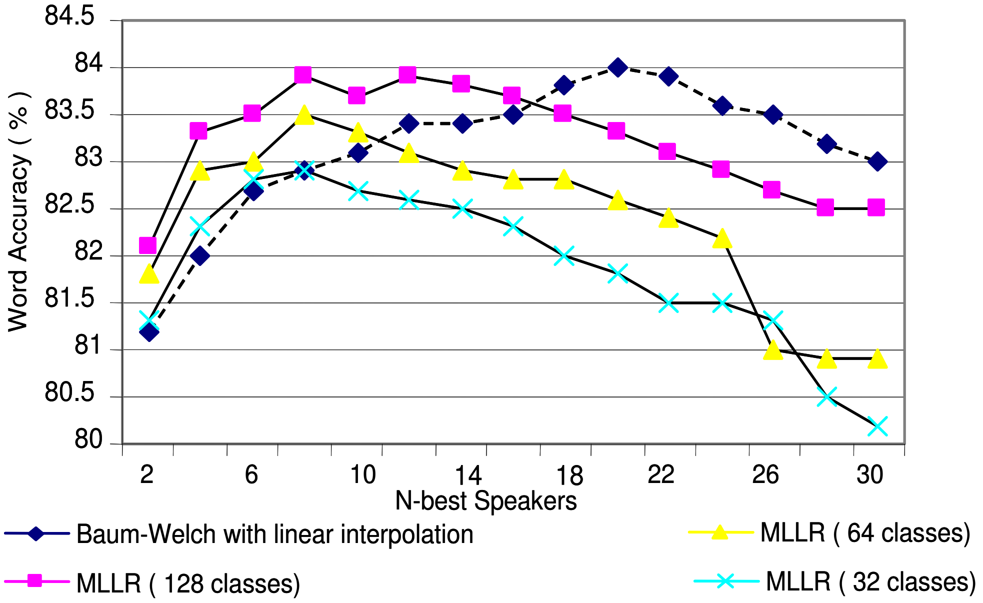


Figure 79. Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Senior Male).

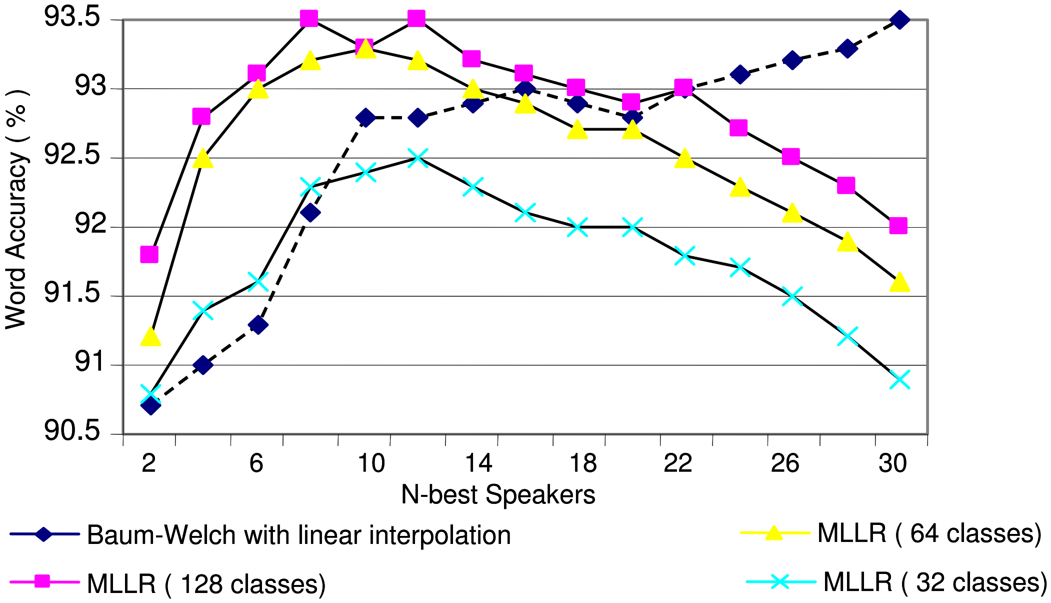


Figure 80. Performance of the proposed rapid MLLR adaptation based HMM-Sufficient Statistics (Senior Female).

difference increases as well.

7.6 Further Investigating the HMM-Sufficient Statistics

We investigate the effects brought by implementing the Baum-Welch based linear interpolation and the online MLLR approach in HMM-Sufficient Statistics adaptation. In Figure 81 we show the graph of the HMM-Sufficient Statistics, particularly the mixture component occupancy (in logscale) versus the pool of all Gaussian mixtures. In this figure, we show the effect of merely reducing N-best from 40 to 25 (without interpolation). This is manifested by a decrease in the mixture component occupancy as depicted by the shifting of the envelope (N-best=25) downwards relative to N-best=40. This can be translated to a reduction in the recognition performance, because reducing the number of selected N-best means reducing the adaptation data. On the other hand, the effect of the linear interpolation pushes back the envelope of the N-best=25 close to N-best=40. The supposed decrease in the mixture component occupancy is compensated by

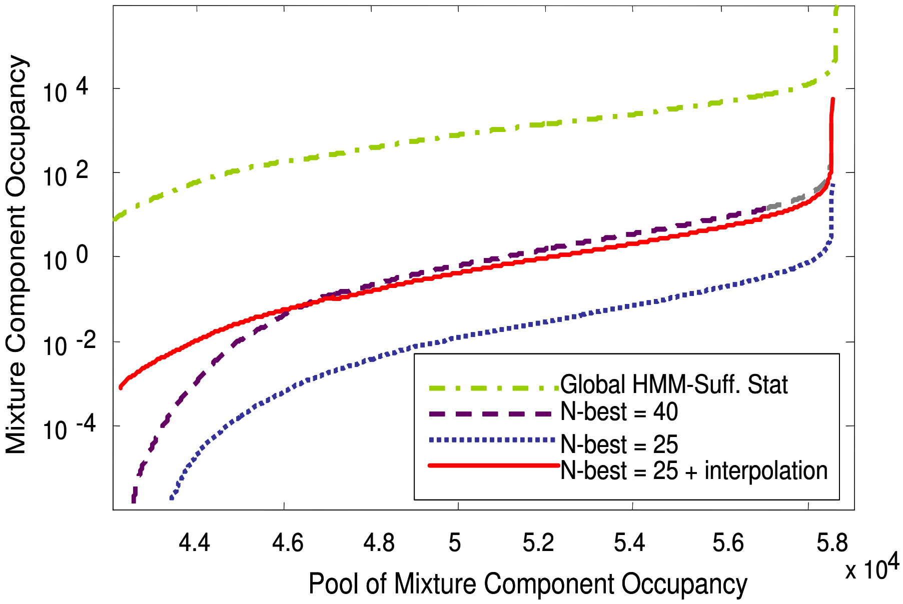


Figure 81. Effects of linear interpolation of the global HMM-Sufficient Statistics.

the interpolation of the global HMM-Sufficient Statistics . This would mask the detrimental effect in the recognition performance brought by decreasing N-best.

Figure 82 illustrates the advantage of using MLLR instead of Baum-Welch. The ordinate shows the number of unupdated models while the abscissa represents the number of N-best selected speakers. This figure shows that the number of unupdated models decreases as N-best increases. Meaning, Baum-Welch based approach is not robust when using only very few N-best. Thus the system can benefit much more when using MLLR.

7.7 Comparisons with VTLN, MAP and Conventional MLLR

We refer to MLLR as the conventional approach requiring speech utterances for adaptation not unless specified as online MLLR using only one arbitrary utterance for selecting N-best speakers' HMM-Sufficient Statistics. Recognition experiments using VTLN, MLLR, MAP were performed for comparison with the HMM-Sufficient Statistics techniques. We also combined VTLN with MLLR (VTLN+MLLR) and VTLN with MAP (VTLN+MAP) for an improved perfor-

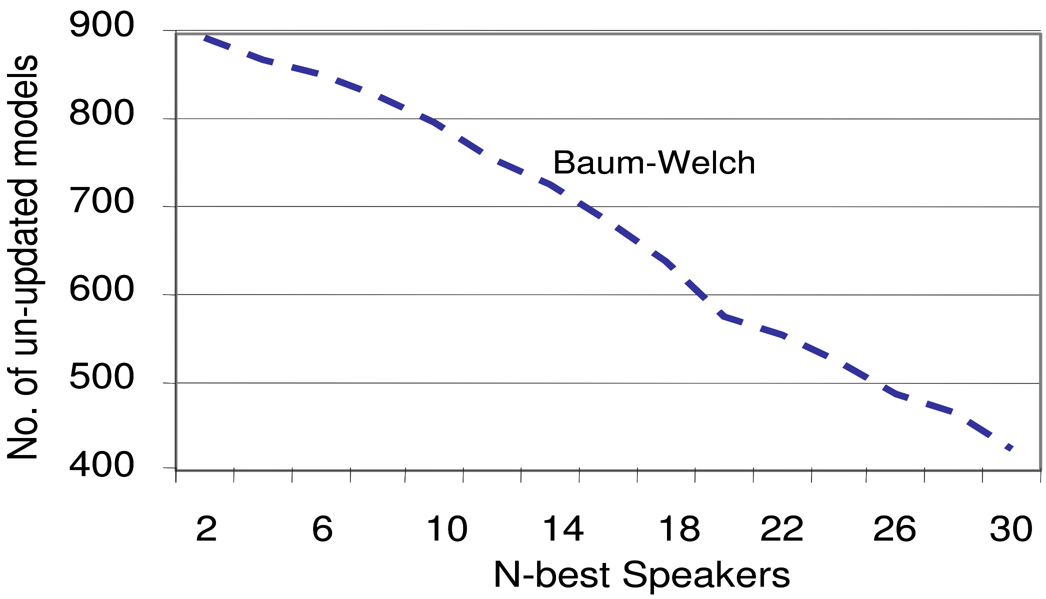


Figure 82. unupdated models caused by insufficiency of adaptation data in Baum-Welch reestimation.

mance and compare it with HMM-Sufficient Statistics adaptation. Figure 83 shows the case of combining VTLN and MAP/MLLR. In the *offline* part of this figure, we search for the warping parameter α that maximizes the log-likelihood score of the training database [49]. Figure 84 shows a plot of α averaged with all speakers, the corresponding α that results to the peaking of the envelope is chosen in warping all of the training utterances and used to reestimate the VTLN-adapted model. Consequently, in the *online* part, we do the same process of finding α of the adaptation utterances using the VTLN-adapted model and warped these utterances prior to MLLR/MAP adaptation. The process of finding α is repeated again for the last time using the MLLR/MAP adapted model to the test utterances. Finally, we warp the testing utterances for recognition experiment.

In Figure 85, we show the recognition results using the supervised MAP, MLLR, VTLN+MAP and VTLN+MLLR. In the abscissa, the labels 10 and 50 utterances correspond to the adaptation data matched with the test speaker for the MLLR and MAP variants. We compare these results with the best per-

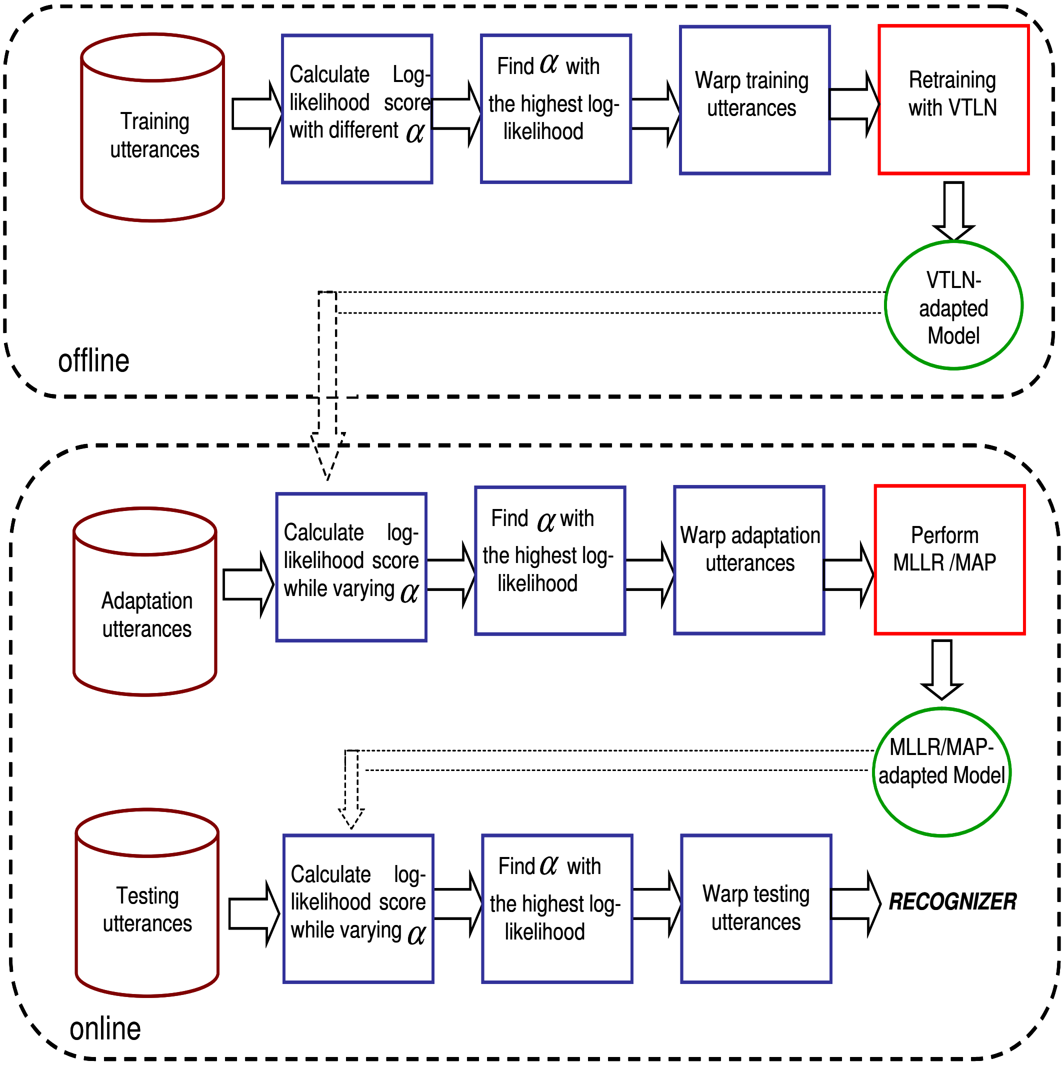

Training utterances
Calculate Log-likelihood score with different α
Find α with the highest log-likelihood
Warp training utterances
Retraining with VTLN
VTLN-adapted Model
offline
Adaptation utterances
Calculate log-likelihood score while varying α
Find α with the highest log-likelihood
Warp adaptation utterances
Perform MLLR /MAP
MLLR/MAP-adapted Model
Testing utterances
Calculate log-likelihood score while varying α
Find α with the highest log-likelihood
Warp testing utterances
RECOGNIZER
online

Figure 83. Block diagram of the supervised VTLN adaptation in finding for the optimum α.

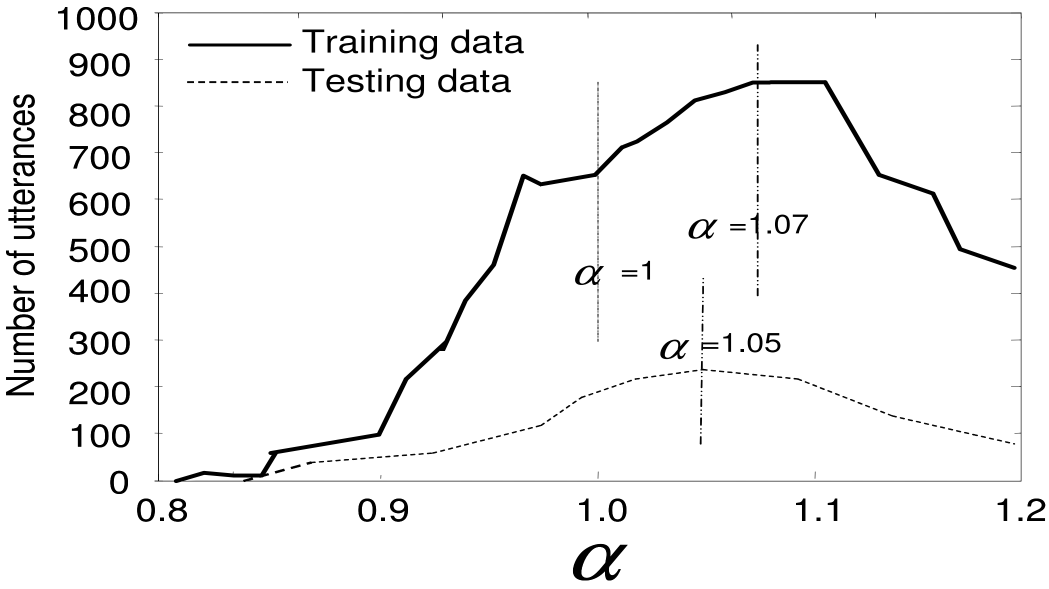


Figure 84. Average values of α used in VTLN adaptation.

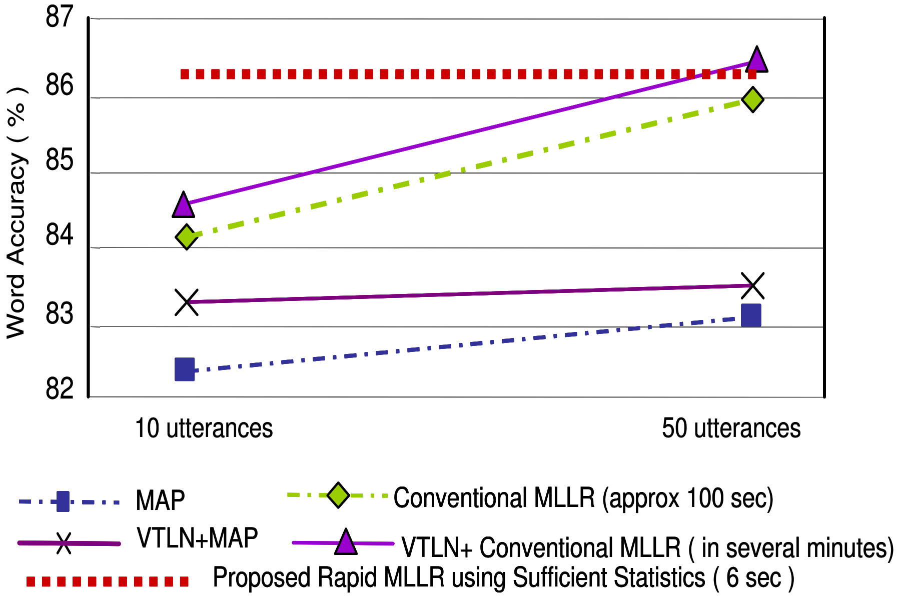
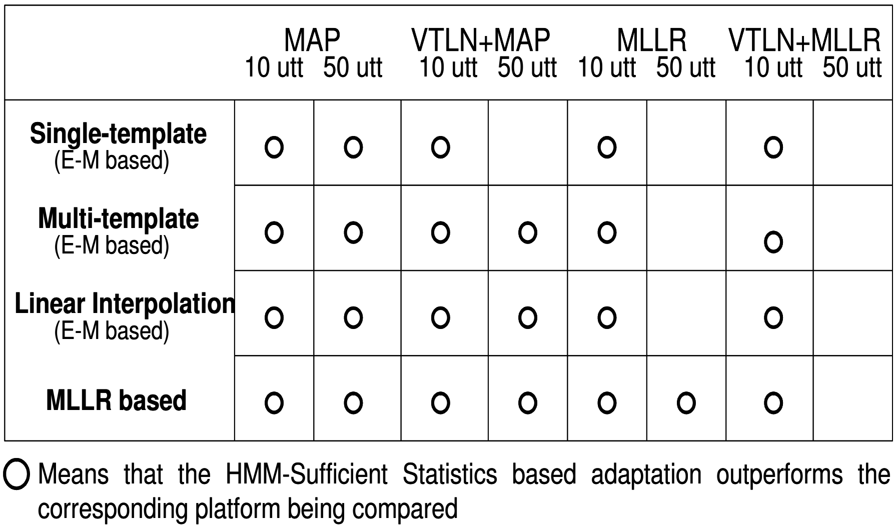


Figure 85. Recognition performance with various adaptation techniques.



	MAP		VTLN+MAP		MLLR		VTLN+MLLR	
	10 utt	50 utt	10 utt	50 utt	10 utt	50 utt	10 utt	50 utt
Single-template (E-M based)	○	○	○		○		○	
Multi-template (E-M based)	○	○	○	○	○		○	
Linear Interpolation (E-M based)	○	○	○	○	○		○	
MLLR based	○	○	○	○	○	○	○	

○ Means that the HMM-Sufficient Statistics based adaptation outperforms the corresponding platform being compared

Figure 86. Performance comparison of HMM-Sufficient Statistics vs. VTLN, MAP, and MLLR.

forming Baum-Welch based (using interpolation) and the online MLLR HMM-Sufficient Statistics (128-class). The abscissa denotes the number of adaptation utterances used. The horizontal block line refers to the Baum-Welch based HMM-Sufficient Statistics with interpolation while the broken line is of the online MLLR based on HMM-Sufficient Statistics. Both of these require only one arbitrary adaptation utterance. This figure shows that both the Baum-Welch and the online MLLR based on HMM-Sufficient Statistics outperform the supervised MLLR, MAP, VTLN+MAP and VTLN+MLLR when using 10-utterance adaptation data. With 50 utterances of adaptation data, MLLR and VTLN+MLLR perform better than the Baum-Welch HMM-Sufficient Statistics . However, the online MLLR based on HMM-Sufficient Statistics is better than conventional MLLR with 50 adaptation data. It should be noted that when using 50-utterances of adaptation data, MLLR and MAP takes more adaptation time than Baum-Welch with interpolation and the online MLLR based on HMM-Sufficient Statistics which can adapt in just 6 and 7 sec respectively using only a single arbitrary adaptation utterance without transcriptions. Figure 86 shows the summary of the comparisons among VTLN, MAP and MLLR. The legend *o* denotes that the corresponding HMM-Sufficient Statistics adaptation outperforms the corresponding adaptation approach being compared with.

7.8 Summary

In this chapter, we presented the set-up used in the experiment. The results of the HMM-Sufficient Statistics adaptation discussed in Chapters 4 and 5 are summarized here showing the evolution of the proposed rapid unsupervised speaker adaptation based on HMM-Sufficient Statistics. Moreover, experiments with the commonly used adaptation techniques in speaker adaptation such as VTLN, MAP, conventional MLLR are compared with our approach. Furthermore, detailed explanation of the merits of using the proposed technique is presented.

8. Evaluation in Real Environment Conditions

The ultimate goal of the rapid unsupervised speaker adaptation is to realize an ASR system with the recognizer and adaptation algorithm working together toward a more robust system. The task of the adaptation algorithm is to provide a robust model to the recognizer using only the current utterance. As a result, the recognizer will have an improved performance for every recognition task. In this chapter, we will discuss the integration of the proposed rapid adaptation in a dialogue system. Furthermore we discuss the modifications of the system's design to accommodate practical integration issues. Lastly we will show recognition results with the current system set-up.

8.1 Practical Implementation

We have shown the potential of the rapid unsupervised adaptation in Chapters 5 and 6. In those chapters, we are more concerned of the performance of the adaptation scheme and the parameters where specific to test the adaptation technique with lesser regards on other issues. This is depicted in Figure 87 where we treat independently every process during the experimental phase to check its effectiveness without the intervention of some other processes. In this figure, we set specific rules or parameters to hasten up testing of the adaptation phase given the adaptation environment. However, this is not the same case in a full system integration. In an actual system implementation as shown in Figure 88 the individual processes shown in Figure 87 are integrated to create a single system, and the focus is not just on the adaptation algorithm alone but in the whole system itself having several processes. This scenario is totally different with that of the experimental phase. Complications arise as more and more processes are involved. These processes might be interdependent with each other or the individual processes may require simultaneous use of the computers' resources. These are just one of the many issues in a full system integration. In Figure 88 for example, both the recognizer and the adaptation module access some system resources to carry out respective tasks. Also, the recognizer needs as input the output adapted model from the adaptation technique. To accommodate these factors affecting the overall system, we need to loosen some of the parameters specifically

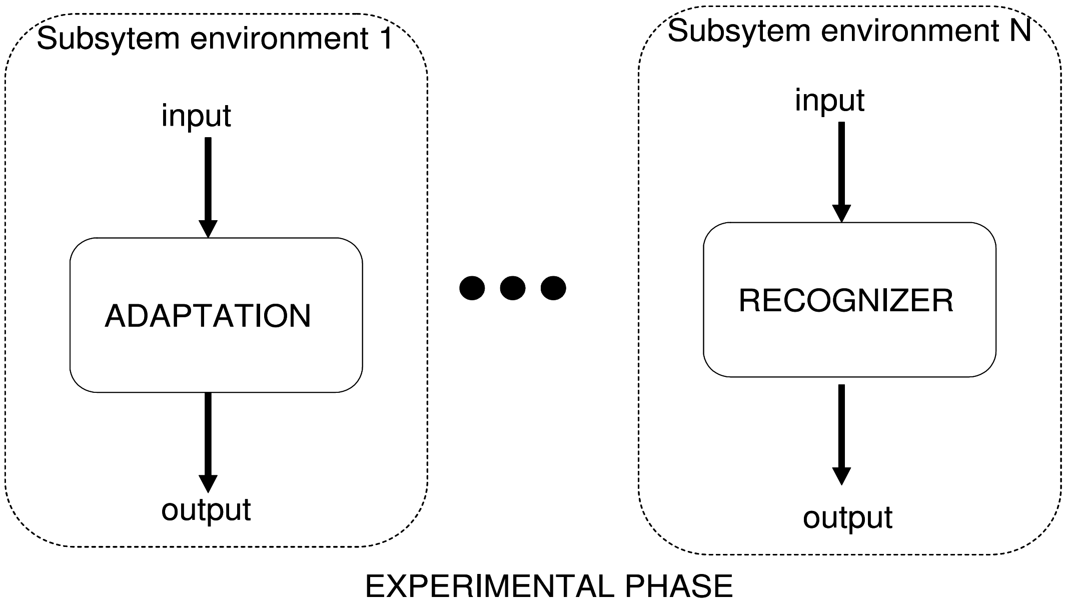


Figure 87. Treating individual system independently to hasten up testing during experimental phase.

designed to each of the processes and create a single environment that supports the interactions of these processes.

8.2 Redesigning The Rapid Adaptation System

The detailed set-up in testing the proposed adaptation system during the experimental phase is shown in Figure 89. In this figure, we already know beforehand the test sets. This means that we have a prior knowledge of which speaker does a certain test utterance comes from. However, we emphasize that we use open test sets which means that these speakers/utterances have never been used in training or in adaptation. To hasten up the evaluation process, we assume that all the test utterances of the same speaker have the same N-best speakers list and every time a test utterance falls to a corresponding speaker, we automatically use the corresponding speaker's N-best list. The purpose of doing this is to avoid calculating the likelihoods each time for a known speaker-utterance and hasten up the evaluation of the recognition performance.

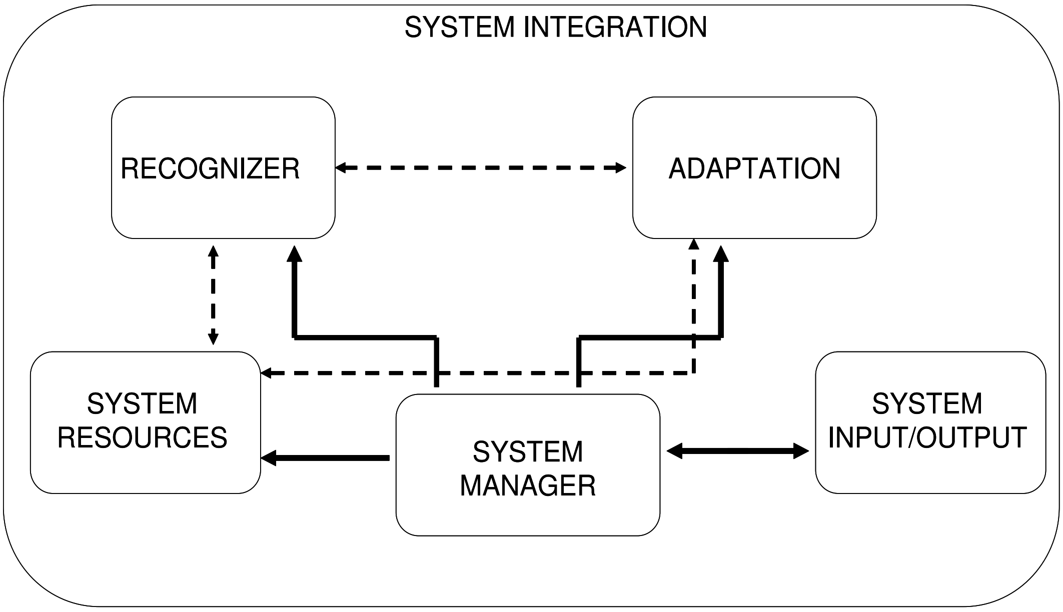


Figure 88. Multiple processes working interdependently in an integrated system.

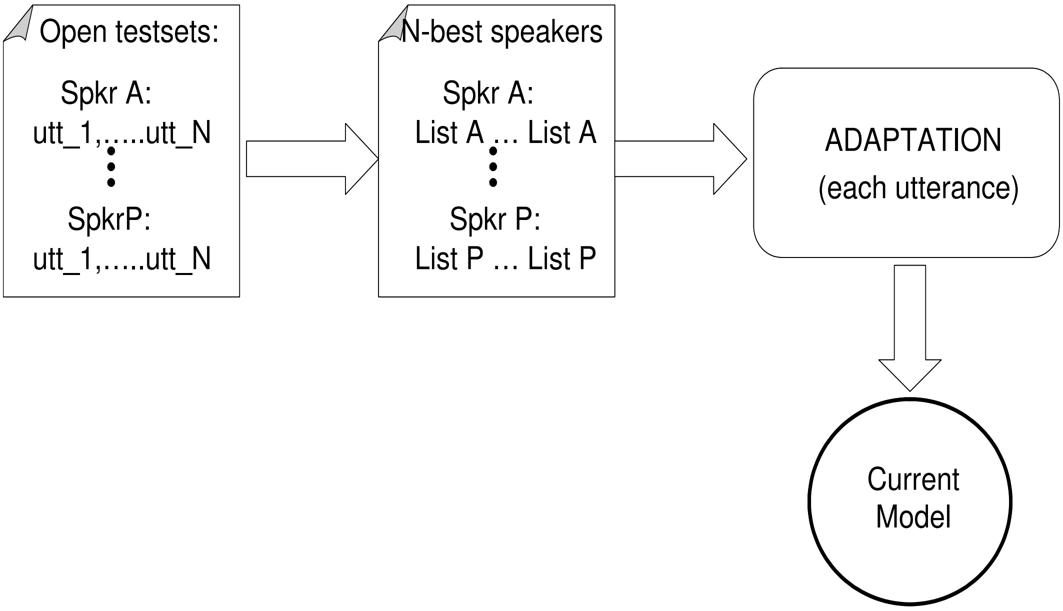


Figure 89. Testing set-up of the rapid unsupervised speaker adaptation (experimental approach).

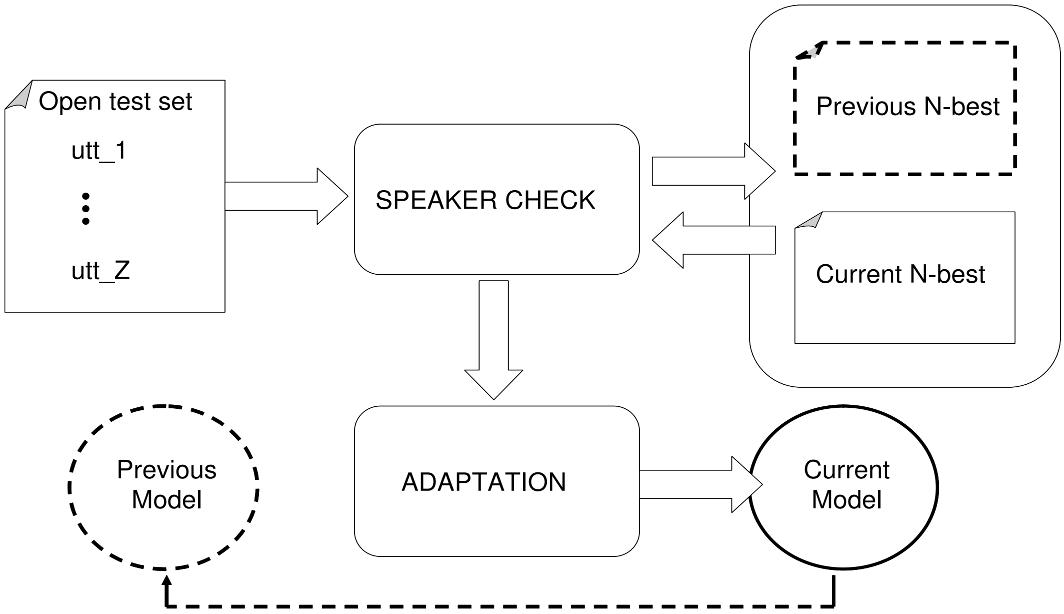


Figure 90. Testing set-up of the rapid unsupervised speaker adaptation (practical implementation).

In a practical implementation, there is no way of identifying the speaker, simply because we do not have prior knowledge of the actual test speakers. It is also impossible to test random speakers which are not in the open test sets because there is no way of comparing results from random speakers reasonably. The only feasible thing to do is to use the same test sets blindly simulating random speakers and at the same time be able to compare results. In Figure 90 the modified implementation of the proposed rapid speaker adaptation is illustrated. In this figure, the utterances in the open test sets are not labeled anymore nor classified to a particular speaker. This means that utterances are shuffled and speaker label removed to simulate an actual test speaker. In real environment scenario, it is more logical to assume that the same speaker may use the system more than once, this means that the next utterance to be processed by the recognizer is more likely to belong to the previous speaker. The speaker adaptation system should be able to exploit these instances when it happens, thus we include a speaker checking mechanism to do this task. This will provide the computation for the individual likelihoods in the speaker-GMMs to search for the N-best speakers list. If the

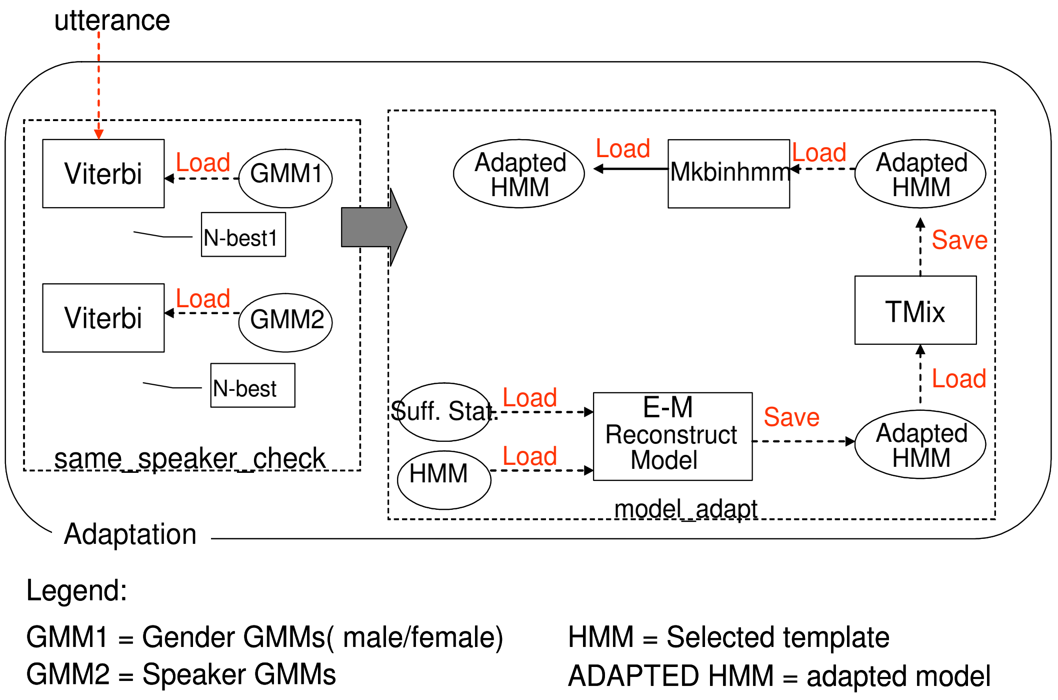


Figure 91. Implementation of the modified adaptation algorithm to accommodate practical issues.

current utterance belongs to the previous speaker then adaptation will not be performed. Otherwise, if the current utterance is recognized to be coming from different speaker, then a new N-best list is created enabling to perform adaptation with a new model. Each time a new speaker is identified, the current N-best list replaces the previous one, the same rule applies to the adapted model.

A very detailed illustration of the modified adaptation subsystem discussed in Figure 90 is shown in the block diagram in Figure 91. In this figure, the unknown test utterance input is processed in a module called, "same speaker check". This module runs the Viterbi algorithm using GMM1 which is gender-GMMs. The corresponding gender will dictate which GMM2 to use. A "male" result, would load the GMM2 composed of male speakers only and a "female" gender result, will load the GMMs composed of female speakers for GMM2. The second Viterbi will find the N-best speakers' list through the likelihoods generated. This module will output gender information and N-best speakers' list in preparation for the adaptation process particularly the selection of the HMM-Sufficient Statistics and the corresponding template model. When all the necessary input to the Baum-Welch adaptation is loaded, rapid adaptation will commence resulting to an adapted model. TMix is then applied which is needed for the PTM models and then transform these into binary format HMMs. In the event that same speaker is detected, the "model adapt" module will not be executed thus the previously adapted model will be considered as current adapted model until such time a newly adapted model is created.

8.3 Performance of the Modified System

We test the recognition performance for the modified online adaptation approach shown in Figure 91 and compare it to the experimental approach. In this testing, we re-shuffled the ordering of the test utterances 5 times as shown in the x-axis. By rearranging the order of the test utterances in random without knowing the speaker, we simulate an actual scenario where actual test speakers are unknown, unlike the "experimental approach" where we know beforehand the test speakers. The recognition performance result for the Baum-Welch and the rapid MLLR implementation using the experimental and the practical approach is shown in Figure 92. In all the five re-shuffles, the changes between the experimental and the practical implementation are insignificant in both the Baum-Welch and MLLR implementation. This signifies that the proposed approach is robust even in practical condition. In this figure the baseline recognition performance without adaptation is also shown with 84.1% WA. It should be noted that the acceptable WA performance of the information guidance task that we use is about 80.0%-84.0%, thus the result of the proposed adaptation algorithm as shown in Figure 92 is considered to be acceptable in information guidance application.

8.4 Implementation With The On-site Dialogue System

As what we have mentioned earlier, the goal of this research is not just to design a rapid adaptation technique, but to implement this technique together with the

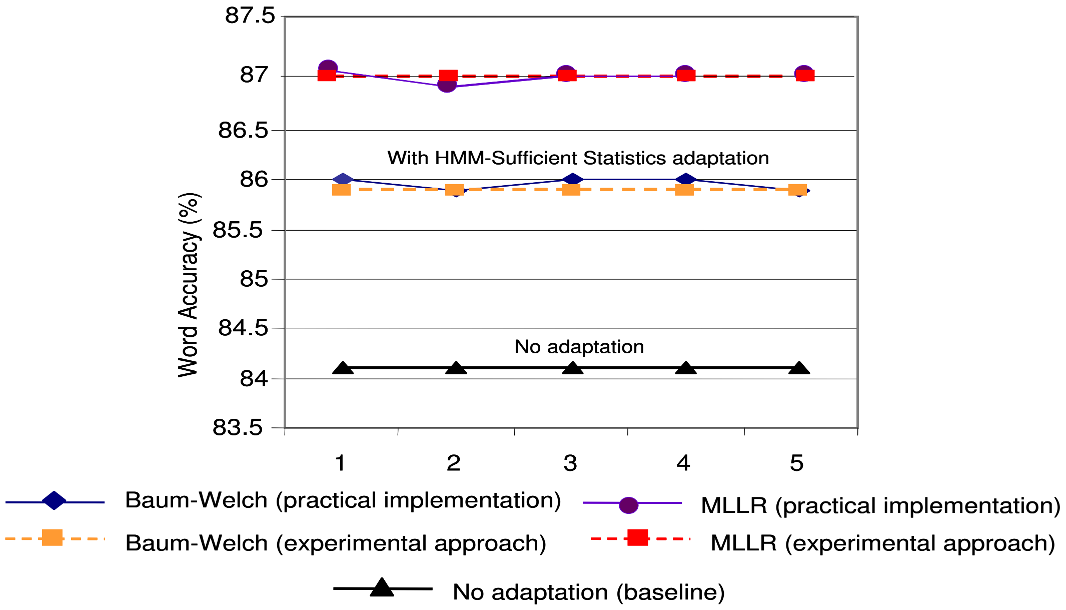


Figure 92. Recognition performance of the practical implementation of the rapid unsupervised speaker adaptation based on HMM-Sufficient Statistics.

recognizer. In this section, we will discuss the actual integration of the rapid adaptation technique in a a dialogue system "Takemaru-kun" shown in Figure 1. This is a guidance dialogue system [3] using automatic speech recognition. A user can address his queries to the system using the microphone without any keyboard and the system will respond accordingly to the queries and give some answers according to its task. In this kind of application, the need to integrate a speaker adaptation scheme arises from the fact that we expect a wide spectrum of users. In addition, this kind of application requires a fast implementation and the only adaptation data that is available to the system is the utterance from the user. In this scenario, the proposed rapid adaptation is more suited to do the job.

Figure 93 shows the dialogue system with the adaptation module interconnected. The dialogue manager is responsible of controlling the flow of all the processes in the whole system. At the beginning, the speech utterance is made available to Julius1, Julius2, and the adaptation module through the audio input tool. In the recognizer side, two processes of "Julius" is used, one uses the fix

SI model (Julius1) and the other uses the adapted model (Julius2). These two recognition processes are controlled via TCP/IP by the dialogue manager. Both of these two processes perform recognition task to the utterance, the dialogue manager then checks which of the two results to a highest likelihood, the hypothesis that corresponds to the highest likelihood is then selected as the recognized words or the final hypothesis. In the adaptation side, the speech utterance is processed independently checking for same speaker and perform adaptation as discussed in Figure 91. Performing the adaptation routine we discussed in the previous chapters and saves the adapted model to a fixed location if adaptation is carried out, otherwise it will leave the fixed location empty to signify that there is no adapted model and that the previous speaker is the owner of the current utterance. The dialogue manager, routinely checks for model in the assigned fixed location. If it finds a model, it automatically sends this model in binary format via TCP/IP to the second recognizer Julius2, the recognizer tasked to process the utterance using the adapted model. As soon as the hypotheses are available from the two recognizers, the results are sent back to the dialogue manager via TCP/IP for the dialogue manager to decide which of the two hypotheses is selected for the actual recognized words. These processes are repeated in a loop until terminated.

It is important to note that in the event when the adaptation process is not yet finish with its current task, and the speaker is already speaking the next utterance, the dialogue system will continue to operate normally and uses the previously adapted model for Julius 2. This event would have a less impact to the system since it is more probable that the current speaker is the previous speaker since abrupt change of users is not common. Besides, the system also uses the SI model and in the event, there is an abrupt change of speaker whose speech is acoustically dissimilar to the previous speaker, the result from the SI model in Julius1 is more likely used.

8.5 Summary

In this chapter, we have extended the implementation design of the proposed rapid adaptation based on HMM-Sufficient Statistics for practical implementation. In Chapter 5 and 6, we dealt about the experimental approach in which

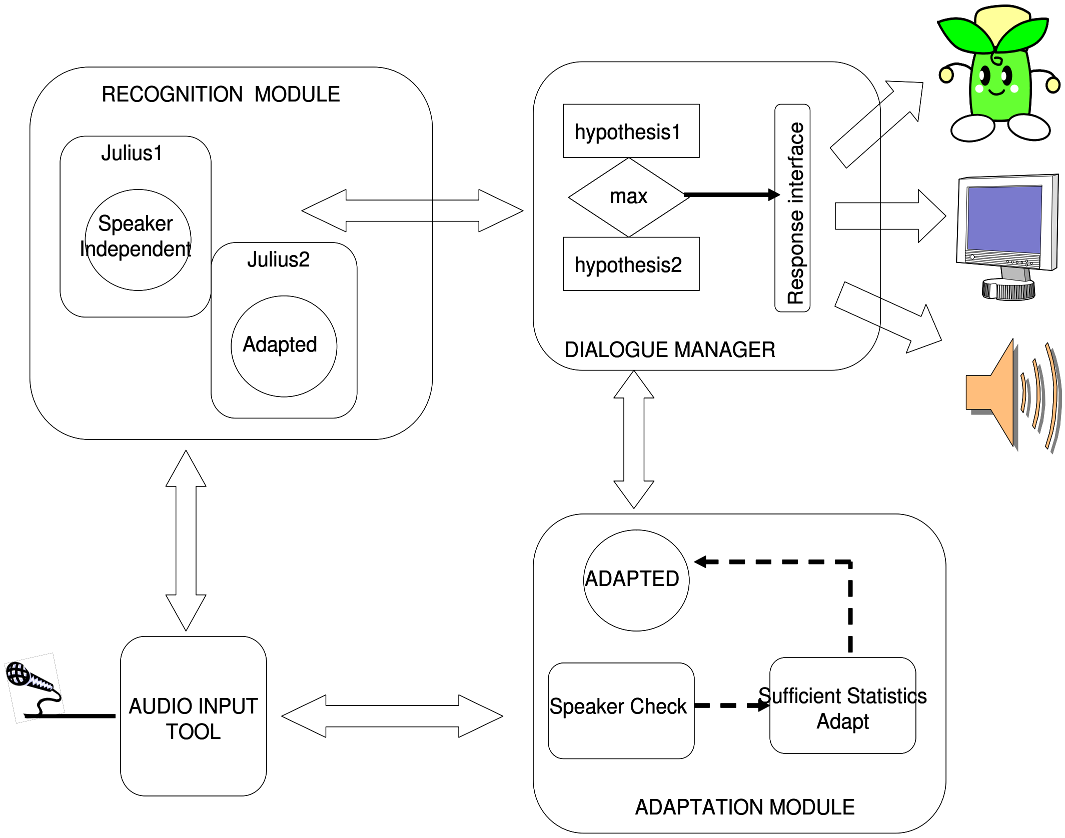


Figure 93. Actual system integration of the adaptation algorithm in the dialogue system.

we focused more on the performance of the adaptation scheme and proved that our approach works well. In this chapter, we considered a real ASR system using the speech dialogue system where the adaptation technique plays a role in the overall processes. We have modified some implementation of the "experimental approach" to work in a real environment situation and our recognition performance evaluation shows that the system is robust and stable even in actual scenarios.

9. Conclusion

In this thesis, we use HMM-Sufficient Statistics as a form of adaptation data instead of using the actual observed data. Since adaptation data can be replaced by the training database using N-best speaker selection, the proposed method breaks the actual E-M process into two separate procedures, the E-step where parameters that are function of the actual observed data are computed in advance and the M-step, where the actual adaptation takes place. In so doing, we can have a rapid unsupervised adaptation using only a single arbitrary utterance.

We have shown two platforms of adaptation schemes where we used HMM-Sufficient Statistics as adaptation data : The Baum-Welch and MLLR techniques. These algorithms supposedly take much more time in executing adaptation when using the actual observed data but we specifically redesigned these two to make use of N-best selected speakers' HMM-Sufficient Statistics as adaptation data. As a result we can have an online adaptation approach in 6 sec adaptation time. Since adaptation is mostly of an offline approach due to the constraints brought by time and adaptation data, this research with rapid adaptation is a manifestation that online adaptation can be achieved in tandem with real-time speech recognition. In addition, by using only a single arbitrary utterance without transcription in carrying out the adaptation process, reinforces the practicability of the proposed method in speaker adaptation. Furthermore, the system works well under office, crowd, booth and car noise and in different SNRs. With the proposed linear interpolation of the HMM-Sufficient Statistics in both the online Baum-Welch and MLLR, it is possible to reduce N-best and a dapt to a robust model. The proposed rapid adaptation which has 87.0% WA performance has an absolute improvement of 2.9% compared to using only the SI model (no adaptation) which has 84.1% of WA. Given the fact that around 80.0% - 84.0% of WA is considered as an acceptable performance in the information guidance task in which we use to test the system, thus an absolute improvement of 2.9% brought by the proposed adaptation algorithm is already considered good-performing technique. Lastly, we have shown that the proposed rapid adaptation technique can be fully integrated in a speech recognition system in Chapter 8. Implementation of the dialogue system together with the recognizer and the adaptation module affirms that the proposed method is feasible in an actual speech recognition application.

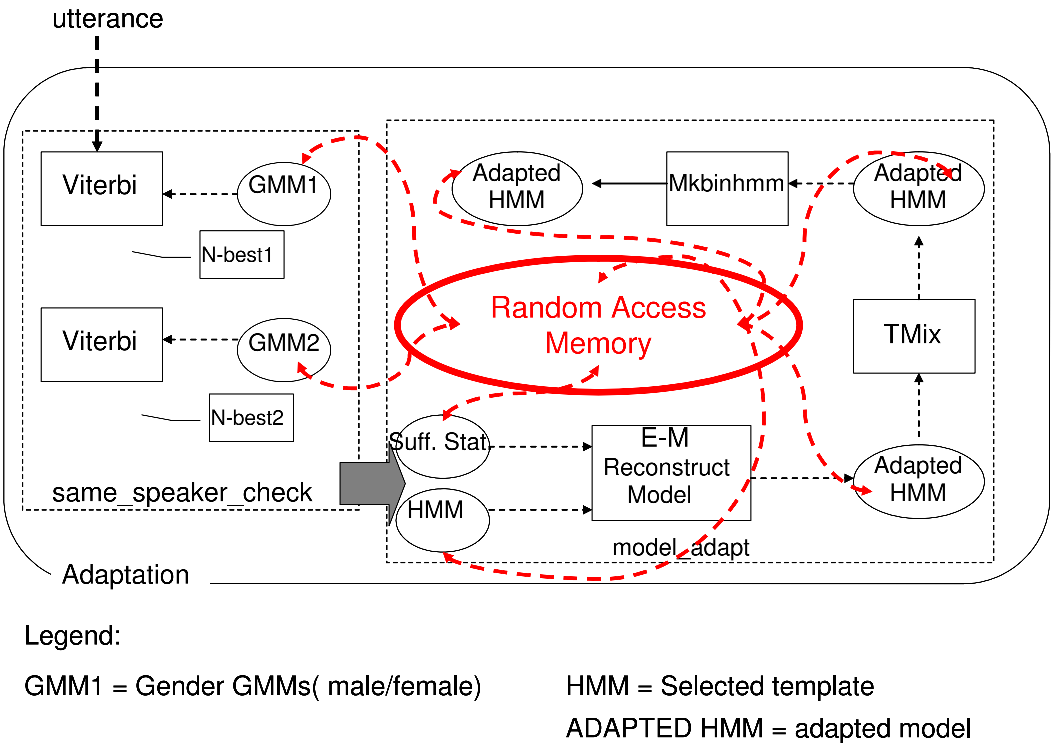


Figure 94. Reducing execution time through effective use of memory in minimizing disk access time in loading and saving parameters.

Moreover, it validates its practicability as far as adaptation time and adaptation data is concerned.

We will focus our future research to make use of existing powerful adaptation techniques to using HMM-Sufficient Statistics for a more improved recognition performance. Investigating discriminative training approach and extend its application to HMM-Sufficient Statistics is an interesting topic. Moreover, we will also look into further detail on how to minimize the execution-time contributed by the adaptation algorithm and the overheads which do not directly involve the adaptation algorithm but affects the overall system in general. This would entail efficient use of memory resources and minimize disk accesses when loading and saving models. As an example, Figure 94 shows an ongoing work on this particular problem. Lastly, we will consider to a great extent the improvements for the N-best speakers' selection. At the moment we are using 64-Mixture GMMs for the individual speaker. We will consider to employ parameter tying to reduce execution time, and most of all find means to improve the N-best performance.

Acknowledgements

First, I would like to thank my research supervisor Prof. Kiyohiro Shikano for giving me the opportunity to join the Speech and Acoustics Laboratory. He has been generous in many ways, his untiring support through discussions, suggestions, and understanding made this thesis possible. His trust provided me the driving force to CREATE. I would also like to thank Asst. Prof. Tomoki Toda for his remarkable contributions to my work. I know he is a very busy man, but he always have more than enough time to sit with me and discuss.

I am grateful to the thesis commitee through Prof. Masatsugu Kidode (co-supervisor) of the Artificial Intelligence Laboratory and Assoc. Prof. Hiroshi Saruwatari for the proof reading and comments. Also, I am thankful to Assoc. Prof. Akinobu Lee for his help in using the methods used in the speech recognizer JULIUS to my research. It is impossible to write the names of all the people who have helped me but i must thank Noriko Watanabe, Toshie Nobori, Shigeki Miyabe, Tobias Cincarek, Yoshimitsu Mori, and the people in the Student Affairs division of NAIST.

My research has been funded by the MEXT e-Society project, this includes funding for my research assistantship and conference trips both domestic and abroad.

Finally, I would like to extend my deepest and heartfelt gratitude to the Japanese government through the Monbukagakusho for the Phd. scholarship. It is an honor to have my doctor's degree conferred in JAPAN.

References

[1] Akinobu Lee, “JULIUS: A Free Continuous Speech Recognition Software” *www.sourceforge.jp*, Nagoya Institute of Technology , Japan

[2] “HTK: Hidden Markov Model Toolkit” *www.htk.eng.cam.ac.uk*, Cambridge University Engineering Department

[3] R. Nishimura, et al., “Takemaru-kun: Speech Oriented Information System for Real World Research Platform” *International Workshop on Language Understanding and Agents for Real World Interaction, pp.70-78, July 2003*, 2003

[4] M. Gales ‘Model-based Techniques for Noise Robust Speech Recognition”, *Doctor’s Dissertation* , Gonville and Caius College, Sept. 1995.

[5] X. Huang, A. Acero, H. Hon “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development”, *Prentice Hall PTR* 2001.

[6] A. Baba, S. Yoshizawa, M. Yamada, A. Lee and K. Shikano, “Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition”, *In Proceedings of EUROSPEECH*, pp. 1657-1660, 2001.

[7] S. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, *IEEE Trans. Acoustic., Speech, Signal Process.*, vol 27, pp 113-120, Apr. 1979.

[8] M. Fujimoto et al. “Large Vocabulary Speech Recognition Under Real Environments Using Adaptive Sub-band Spectral Subtraction”, *In Proceedings of ICSLP*, pp. I-305-308, 2000.

[9] N. Virag, “Speech Enhancement Based on Masking Properties of the Human Auditory System”, *A Master’s Thesis*, Swiss Federal Institute of Technology 2000.

[10] M. Berouti, Schwartz and J. Markhoul, “Enhancement of Speech Corruptedby Acoustic Noise”, *In Proceedings of IEEE Int. Conf. on Acoust., Speech, Signal Procs.*, pp.208-211, April 1979.

[11] H. Saruwatari, S.Kurita, K.Takeda, F. Itakura, T. Nishikawa, and K. Shikano,"Blind Source Separation Combining Independent Component Analysis and Beamforming", *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp 1135-1146, 2003.

[12] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, "Robust Speech Recognition with Spectral Subtraction in Low SNR", *In Proceedings Interspeech*, pp 2077-2080, 2004.

[13] R. Gomez, L. Akinobu, H. Saruwatari, K. Shikano "Wiener Filtering-based Robust Speech Recognition Under Low SNR", *In Proceedings of Acoustical Society of Japan*, Spring Meeting 2004.

[14] D.A. Florencio, H. S. Malvar, "Multi-channel Filtering for Optimum Noise Reduction in Microphone Arrays", *In Proceedings of ICASSP*, May 2001.

[15] A Girardi, I. Sanches, "Multi-dimensional Filtering for Speech Enhancement via Microphone Arrays", *In Proceedings of NAIST COE International Symposium*, Nara, Japan 2001.

[16] Jean-Claude Junqa "Robust Speech Recognition in Embedded Systems and PC applications", *Kluwer Academic Publishers*, 2000.

[17] C. Huang, T. Chen, S. Li and JL. Zhou, "Analysis of Speaker Variability", *In Proceedings of Eurospeech*,Vol. 2, pp 1377-1380 September 2001.

[18] B. Xiang, L. Nguyen, S. Matsoukas and R. Schwartz, "Cluster-Dependent Acoustic Modeling", *In Proceedings of ICASSP*,Vol.1, pp. 677-680 2005.

[19] D. Pye and P.C. Woodland "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Adaptation", *In Proceedings of ICASSP*,Vol.2,No.1,pp.1047-1051, Apr. 1997.

[20] P. Zhan, M. Westphal, M. Finke and A. Waibel "Speaker Normalization and Speaker Adaptation- A combination for conversational Speech Recognition", *In Proceedings of Eurospeech*, vol. pp. 2087-2090 September 1997.

[21] Giuliani and M. Gerosa, “Investigating Recognition of Children‘s Speech”, *In Proceedings of ICASSP*,Vol 2, pp. 137-140 April 2003.

[22] C.J.Leggeter and Woodland “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, *In Proceedings of Computer Speech and Language*, vol.9,pp.171-185, 1995.

[23] J. Gauvain and C. H. Lee “Maximum A Posteriori Estimation For Multivariate Gaussian Mixture Observation of Markov Chains”, *In IEEE Transactions SAP*, vol 2, pp 291-298, 1994.

[24] C. Huang, T. Chen and E. Chan, “Transformation and Combination of Hidden Markov Models for Speaker Selection Training”, *In Proceedings of ICSLP*, 2004.

[25] T. Matsui, T. Matsouka and S. Furui, “Smoothed N-Best Based Speaker Adaptation for Speech Recognition ”, *In Proceedings of ICASSP*, pp. 1015-1018, 1997.

[26] G. Vatbhava, V. Karthik and G. Ramesh, “Rapid Adaptation with Linear Combinations of Rank-one Matrices”, *In Proceedings of ICASSP*, 2001.

[27] R. Kuhn, F. Perronnin, P. Nguyen, J. Junqua and L. Rigazio, “Very Fast Adaptation with a Compact Context-Dependent Eigenvoice Model”, *In Proceedings of ICASSP*, 2001.

[28] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada and K. Shikano, “Unsupervised Speaker Adaptation Based on Sufficent HMM Statistics of Selected Speakers”, *In Proceedings of ICASSP*, 2001.

[29] Sean Borman ‘The Expectation Maximization Algorithm” *A Short Tutorial* 2004

[30] Frank Dellaert ‘The Expectation Maximization Algorithm” *College of Computing, Georgia Institute of Technology Technical Report number GIT-GVU-02-20* 2002

[31] A. Dempster, N. Laird, D. Rubin 'Maximum Likelihood from Incomplete Data via the EM algorithm" *In Proceedings, Journal of the Royal Statistical Society, Series B* 39(1), pp.1-38, 1977

[32] H. Hartley 'Maximum Likelihood Estimation from Inclomplete Data" *Biometrics* pp. 174-194, 1958

[33] G. McLachlan, T. Krishnan 'The EM alorithm and Extensions" *Wiley Series in Probability and Statistics* John Wiley and Sons 1997

[34] R. Neal, G. Hinton 'A View of the EM algorithm that justifies incremental, sparse, and other variants." *Learning in Graphical Models* Kluwer Academic Press 1998

[35] M. Tanner, 'Tools For Statistical Inference" Springer Verlag, New York. 3rd ed.

[36] C. Bishop 'Neural Networks for Pattern Recognition" Oxford: Clarendon Press 1995

[37] W. Buntine 'Computation with the exponential Family and Graphical Models" *Tutotial: NATO Workshop on Learning in Graphical Models* 1996

[38] S. Luttrell 'Partitioned Mixture Distribution: An Adaptive Bayesian Network for Low Level Image Processing" *In Proceedings IEEE on Vision, Image and Signal Processing* pp. 251-260, August 1994

[39] R. Redner, H. Walker 'Mixture Densities, Maximum Likelihood and EM Algorithm" *SIAM Review*pp. 195-239, Aprill 1984

[40] M. Bazaraa, C. Shetty 'Nonlinear Programming" John Wiley and Sons, New York 1979.

[41] T. Minka 'Expectation-Maximization as lower bound maximization" *http://www-white.media.mit.edu/ tp-minka/papers/em.html* 1998.

[42] R. Duda, P. Hart, D. Stork "Pattern Classification" *by John Wiley and Sons*, 2nd Edition, ISBN 0-471-05669-3 2001.

[43] C. Tobias, T. Toda, H. Saruwatari, and K. Shikano 'Utterance-Based Selective Training for the Automatic Creation of Task-Dependent ACoustic Models" *IEICE Special Issue on Statistical Modelling for Speech Processing*, Vol. E89-D, No. 3 March 2006.

[44] S. Yamade, K. Matsunami, A. Baba, A. Lee, H. Saruwatari and K. Shikano , "Spectral Subtraction In Noisy Environments Applied To Speaker Adaptation Based on HMM Sufficient Statistics", *In Proceedings of ICSLP* , pp I-1045-1048 2000.

[45] R. Gomez, A. Lee, H. Saruwatari and K. Shikano "Rapid Unsupervised Speaker Adaptation Based on Multi-template HMM Sufficient Statistics in Noisy Environments", *In Proceedings of EUROSPEECH*, pp 296-301, 2005.

[46] R. Gomez, T. Toda, H. Saruwatari and K. Shikano, "Improving Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics", *In Proceedings of ICASSP*, pp 1001-1004, 2006.

[47] R. Gomez, A. Lee, H. Saruwatari and K. Shikano "Speaker-Class Reduction for HMM-Sufficient Statistics Adaptation Using Multiple Acoustic Models", *In Proceedings of Acoustical Society of Japan*, March 2005.

[48] A. Lee, T. Kawahara, K. Takeda and K. Shikano, "A New Phonetic Tied-Mixture Model For Efficient Decoding", *In Proceedings of ICASSP* , pp. 1269-1272 2000.

[49] R. Gomez, A. Lee, T. Toda, H. Saruwatari and K. Shikano, "Improving Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics In Noisy Environments Using Multi-template Models", *IEICE Special Issue on Statistical Modelling for Speech Processing*, Vol. E89-D, No. 3 March 2006.

Appendix

A. List of Abbreviations

ASR Automatic Speech Recognition
dB decibel
ETSI European Telecommuniactions Standard Institute
E-M Expectation Maximization
GMM Gaussian Mixture Model
HMM Hidden Markov Model
JNAS Japanese Newspaper Article Sentences
MAP Maximum A-Priori
MLLR Maximum Likelihood Linear Regression
MelCD Mel Cepstrum Distortion
MFCC Mel Frequency Cepstrum Coefficients
NRR Noise Reduction Rate
SS Spectral Subtraction
SI Speaker Independent
SNR Signal-to-Noise Ratio
WA Word Accuracy
VTLN Vocal Tract Length Normalization

B. Expressions Used in Weighting the Sufficient Statistics

The new expression for the single-template Baum-Welch adapted HMM-Sufficient Statistics previously given in Equations 64-67 which currently reflects the weights w_s are given as follows,

$$
\begin{aligned}
C_{im}^{adp} &= \frac{\sum_{s=1}^{S} w_s L_{im}^{s}}{\sum_{s=1}^{S} \sum_{m=1}^{M} w_s L_{im}^{s}}, && (103)\\
\boldsymbol{\mu}_{im}^{adp} &= \frac{\sum_{s=1}^{S} w_s \boldsymbol{m}_{im}^{s}}{\sum_{s=1}^{S} w_s L_{im}^{s}}, && (104)\\
\boldsymbol{\Sigma}_{im}^{adp} &= \frac{\sum_{s=1}^{S} w_s \boldsymbol{v}_{im}^{s}}{\sum_{s=1}^{S} w_s L_{im}^{s}} - \boldsymbol{\mu}_{im}^{adp} \boldsymbol{\mu}_{im}^{adp\,T}, && (105)\\
a_{ij}^{adp} &= \frac{\sum_{s=1}^{S} w_s L_{ij}^{s}}{\sum_{s=1}^{S} \sum_{j=1}^{J} w_s L_{ij}^{s}}. && (106)
\end{aligned}
$$

Publications

A. Journals

[1] Randy Gomez, Akinobu Lee, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Reducing Computation Time of the Rapid Unsupervised Speaker Adaptation based on HMM-Sufficient Statistics” (Accepted : August 7, 2006)
[2] Mitsuru Samejima, Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano,”Evaluation of Acoustics Model and Unsupervised Speaker Adaptation for Child Speech Recognition in Real Environment”vol.47, 2006
[3] Randy Gomez, Akinobu Lee, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Improving Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics In Noisy Environments Using Multi-template Models”, IEICE Special Issue on Statistical Modelling for Speech Processing , Vol. E89-D, No. 3, pp. 998-1105, March 2006

B. International Conferences

[1] Randy Gomez, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Improving Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics”, In Proceedings of ICASSP, pp. 1001-1004, September 2006
[2] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano “Rapid Unsupervised Speaker Adaptation Based on Multi-template HMM Sufficient Statistics in Noisy Environments”, In Proceedings EUROSPEECH, pp. 296-301, September 2005.
[3] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “Robust Speech Recognition with Spectral Subtraction in Low SNR”, In Proceedings Interspeech, pp. 2077-2080, March 2004

C. Meetings

[1] Randy Gomez, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Improving Rapid MLLR-based Unsupervised Speaker Adaptation using HMM-Sufficient

Statistics”, To be published in the Acoustical Society of Japan, Autumn Meeting, September 2006

[2] Randy Gomez, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Fast Implementation of Maximum Likelihood Linear Regression for Unsupervised Speaker Adaptation using HMM-Sufficient Statistics”, Acoustical Society of Japan, Spring Meeting, pp. 65-66, March 2006

[3] Randy Gomez, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Improving the Rapid Unsupervised Speaker Adaptation Through HMM-Sufficient Statistics Weighting”, Acoustical Society of Japan, Spring Meeting, pp. 155-156, March 2006

[4] Randy Gomez, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano,”Evaluating Rapid Unsupervised Speaker Adaptation Using Linear Interpolation of HMM-Sufficient Statistics”, SLP, pp. 13-18, December 2005

[5] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “Speaker-Class Reduction for HMM Sufficent Statistics Adaptation using Multiple Acoustic Models”, Acoustical Society of Japan, Autumn Meeting, pp. 133-134, September 2005

[6] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “Unsupervised Speaker Adaptation Based on HMM Sufficient Statistics Using Multiple Acoustic Models Under Noisy Environment”, SLP, pp. 205-210, December 2005

[7] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “Rapid Multi-template HMM Sufficient Statistics Unsupervised Speaker Adaptation with Linear Interpolation”, Acoustical Society of Japan, Spring Meeting, pp. 111-112, March 2005

[8] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “ Effects of Multiple Acoustic Models in Unsupervised Speaker Adaptation Based on HMM Sufficient Statistics”, Acoustical Society of Japan, Autumn Meeting, pp. 179-180, September 2004

[9] Randy Gomez, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, “Wiener Filtering-Based Robust Speech Recognition Under Low SNR”, Acoustical Society of Japan, Spring Meeting, pp. 171-172, March 2004