??
3
caseId_SpeakerId_Age_Sex_StartTime_EndTime
??
JNAS
S-
JNAS
4
Feature extraction

Acoustic models
??
??

Language models

Test utterances
younger adult
older adult
younger adult
older adult
??
Language Model Perplexity and OOV rate (%)
***Younger adult* test set** ***Older adult* test set**
younger adult
older adult

Baseline results
younger adult
older adult
??
$p <$ 0.001
??
??
??
younger adult
older adult
??
??
??
??
02, 03, 04, 05
08
07
10
??
??
older adult
??
??
older adult
04, 07
08
$p <$ 0.01
04
08

Design
(LM-Older)
(LM-Young)
(LM-All-1)
(LM-Older-1)
(LM-Young-1)
LM-All-

THE UNIVERSITY OF EDINBURGH

??

$\widehat{\quad}_{w}\, p \mid$

1

$\{ \quad , \quad , \ldots, \ _{T} \}$

$p \mid$

??

$\widehat{\quad}_{w}\, p \mid \qquad _{w} \left\{ \underline{p \quad p \mid \quad} \right\}$

2

p

$p \mid$

??

p

$p \mid$

??

$_{w} \qquad . \, p$

$_{s}p \qquad . \, p$

$_{v} \qquad\quad . \, p$

??

??

$1 - \qquad ^{-} 3$

$0 \leq \quad < 1$

$$\left(0.54 - 0.4 \, \cos \left(\frac{2}{\quad - 1} \right) \right)$$

4

$$_{mel} \quad 1127 \, \text{og} \, _{e} \left(1 \quad \frac{}{700} \right)$$

5

??

$_{j}$

$_{j}$

$$_{i} \quad \sqrt{\frac{2}{}} \sum_{j}^{N} \, \text{og} \quad _{j} \, \cos \left(— \quad - 0.5 \quad \right)$$

$$\frac{1}{1 \quad \sum_{i}^{p} \quad -i}$$

7

$_{i}$

3

??

$_{ij}$

$_{ij}$

$p \ _{j}$

$1 \mid \ _{i}$

$_{i}$

$_{ij}$

$\sum_{j}^{N} \quad _{ij}$

1

$_{j} \quad _{t}$

$_{j} \quad _{t}$

$p \quad \mid \ _{j}$

$$_{j} \quad _{t} \qquad \frac{1}{\sqrt{2 \quad ^{n} \mid \ _{j} \mid}} \ ^{-\frac{1}{2}} \ ^{\varphi - \mu_{j}} \ ^{\prime} \ ^{-1} \ ^{\varphi - \mu_{j}}$$

$_{j}$

$_{j}$

1

Automatic Speech Recognition for ageing voices

Ravichander Vipperla

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2011

Abstract

With ageing, human voices undergo several changes which are typically characterised by increased hoarseness, breathiness, changes in articulatory patterns and slower speaking rate. The focus of this thesis is to understand the impact of ageing on Automatic Speech Recognition (ASR) performance and improve the ASR accuracies for older voices.

Baseline results on three corpora indicate that the word error rates (WER) for older adults are significantly higher than those of younger adults and the decrease in accuracies is higher for males speakers as compared to females.

Acoustic parameters such as jitter and shimmer that measure glottal source disfluencies were found to be significantly higher for older adults. However, the hypothesis that these changes explain the differences in WER for the two age groups is proven incorrect. Experiments with artificial introduction of glottal source disfluencies in speech from younger adults do not display a significant impact on WERs. Changes in fundamental frequency observed quite often in older voices has a marginal impact on ASR accuracies.

Analysis of phoneme errors between younger and older speakers shows a pattern of certain phonemes especially lower vowels getting more affected with ageing. These changes however are seen to vary across speakers. Another factor that is strongly associated with ageing voices is a decrease in the rate of speech. Experiments to analyse the impact of slower speaking rate on ASR accuracies indicate that the insertion errors increase while decoding slower speech with models trained on relatively faster speech.

We then propose a way to characterise speakers in acoustic space based on speaker adaptation transforms and observe that speakers (especially males) can be segregated with reasonable accuracies based on age. Inspired by this, we look at supervised hierarchical acoustic models based on gender and age. Significant improvements in word accuracies are achieved over the baseline results with such models. The idea is then extended to construct unsupervised hierarchical models which also outperform the baseline models by a good margin.

Finally, we hypothesize that the ASR accuracies can be improved by augmenting the adaptation data with speech from acoustically closest speakers. A strategy to select the augmentation speakers is proposed. Experimental results on two corpora indicate that the hypothesis holds true only when the amount of available adaptation is limited to a few seconds. The efficacy of such a speaker selection strategy is analysed for both younger and older adults.

Acknowledgements

First and foremost, I am sincerely grateful to my thesis advisor Prof. Steve Renals for his expert guidance, support and encouragement throughout the period of my doctoral work. His deep understanding of the subject matter has been an invaluable resource for this research work. He has been a wonderful mentor and has inspired me in several ways to pursue scientific quest further in my life.

I am also deeply indebted to my supervisor Dr. Joe Frankel who despite moving on to become an entrepreneur, found time from his busy schedule regularly to review my work and provide helpful guidance.

I would like to express my sincere thanks to Dr. Maria Wolters for providing critical feedback on my work and to Prof. Simon King who has given me valuable advice time and again and for helping me get started with cluster computers.

I am extremely thankful to Prof. Phil Green and Prof. Hiroshi Shimodaira for agreeing to be on my examination panel and for providing me with some constructive feedback to improve this manuscript.

The financial support for this work from Scottish Funding Council and HCRC, University of Edinburgh is gratefully acknowledged. I wish to thank all the members of the MATCH project for providing a nice collaborative research environment and helping me broaden my perspective on a wider range of technologies suited for home care systems.

I am indebted to Dr. Junichi Yamagishi, Dr. Giulia Garau, Dr. Mike Lincoln, and other members of CSTR for always extending a helping hand to resolve issues in experimental design and setup. I would also like to thank Prof. Mark Liberman and Prof. Jerry Goldman for their advice in setting up experiments using the SCOTUS corpus. I have cherished the company of my colleagues at CSTR with whom I have shared unforgettable hours of fun and intellect uplift. I am thankful to them for making this whole experience so much more worthwhile.

I would like to acknowledge the timely support from our wonderful admin and IT assist teams, and the infrastructure provided by the cluster compute team, Edinburgh university library and the School of Informatics.

Several open source tools such as HTK, HTS, Praat, Cluto, LibSVM, and R have been used in this work. I sincerely thank the developers of these tools for their effort.

Finally and most importantly, I owe my loving thanks to my wife Neelima, my parents (Shri. V. Nagendra Rao and Smt. V. Asoka Rani), my sister and my close family,

my extended family and my friends who have put up with me with patience through this roller coaster ride and for being a constant source of support and encouragement.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Ravichander Vipperla*)

To shri Ganapati deva.

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Objectives 2

1.3 Publications 4

2 Ageing voices **5**

2.1 Human speech production mechanism 5

2.2 Changes in the speech production mechanism with ageing 6

2.2.1 Changes in the Respiratory system 7

2.2.2 Changes in the Larynx 8

2.2.3 Changes in the vocal tract 10

2.2.4 Neuromuscular control 11

2.3 Acoustic effects of ageing 12

2.3.1 Average fundamental frequency 12

2.3.2 Fundamental frequency variation and Amplitude variation . . 13

2.3.3 Jitter 13

2.3.4 Shimmer 14

2.3.5 Breathiness 15

2.3.6 Sound pressure level 17

2.3.7 Speech rate 17

3 Automatic Speech Recognition **18**

3.1 ASR architecture 18

3.1.1 Feature extraction 20

3.1.2 Acoustic models 28

3.1.3 Language models 37

3.1.4 Lexicon 43

3.1.5 Decoder . . . 44
3.1.6 Performance measures . . . 45
3.2 Normalisation approaches in acoustic space . . . 46
3.2.1 Cepstral mean and variance normalisation . . . 46
3.2.2 Vocal tract length normalisation . . . 47
3.3 Adaptation approaches in acoustic space . . . 49
3.3.1 Maximum likelihood adaptation . . . 49
3.3.2 Maximum a posteriori adaptation . . . 53
3.3.3 Speaker space adaptation approaches . . . 57
3.4 Automatic age recognition . . . 58
3.5 Automatic speech recognition on older voices . . . 60

4 ASR accuracy on ageing voices: Baseline Experiments 62
4.1 Corpora . . . 62
4.1.1 SCOTUS . . . 63
4.1.2 MATCH . . . 64
4.1.3 JNAS . . . 65
4.2 ASR WERs on older voices . . . 67
4.2.1 Experiments with SCOTUS corpus . . . 67
4.2.2 Experiments with MATCH corpus . . . 73
4.2.3 Experiments with JNAS corpus . . . 76
4.3 Summary . . . 79

5 Impact of changes in glottal source parameters with ageing on ASR 80
5.1 Experimental setup . . . 80
5.2 Fundamental frequency . . . 81
5.3 Jitter . . . 84
5.4 Shimmer . . . 86
5.5 Harmonicity . . . 88
5.6 Summary . . . 89

6 Articulatory changes in older voices 90
6.1 Phoneme recognition accuracies . . . 90
6.1.1 Results on the SCOTUS corpus . . . 92
6.1.2 Longitudinal results on the SCOTUS corpus . . . 92
6.1.3 Results on the JNAS corpus . . . 94

6.2 Vowel centralisation . . . 95
6.3 Speaking rate . . . 98
6.3.1 Speaking rate comparison on SCOTUS corpus . . . 99
6.3.2 Speaking rate comparison on JNAS corpus . . . 99
6.3.3 Impact of speaking rate changes on ASR accuracies . . . 101
6.4 Summary . . . 104

7 Acoustic models for older voices 105
7.1 Speaker classification and clustering . . . 105
7.1.1 Age group classification using SVMs . . . 106
7.1.2 Speaker clustering based on MLLR transforms . . . 107
7.2 Supervised hierarchical models . . . 109
7.2.1 Experimental setup . . . 110
7.2.2 Results . . . 111
7.3 Unsupervised hierarchical models . . . 112
7.3.1 Acoustic models . . . 112
7.3.2 Testing . . . 113
7.3.3 Results . . . 114
7.4 Modifying HMM transition parameters . . . 114
7.4.1 Experimental results on the JNAS corpus . . . 116
7.4.2 Experimental results on the SCOTUS corpus . . . 117
7.4.3 Discussion . . . 117
7.5 Summary . . . 118

8 Speaker selection to augment adaptation data 119
8.1 Distance measure . . . 120
8.2 Speaker identification task . . . 121
8.3 Speaker selection . . . 122
8.4 Experiments . . . 124
8.4.1 Experiments with the AMI corpus . . . 124
8.4.2 Experiments on SCOTUS Corpus . . . 126
8.4.3 Results . . . 128
8.4.4 Discussion . . . 130
8.5 Summary . . . 133

9 Conclusions **134**
9.1 Summary 134
9.2 Future work 137

A Appendix: Experimental result tables **139**

Bibliography **147**

List of Figures

2.1 Human speech production mechanism 6
2.2 Cepstral peak prominence 16

3.1 Parametric representation of speech 19
3.2 Automatic speech recognition system 20
3.3 Speech signal and vocal tract response 22
3.4 MFCC and PLP feature extraction 23
3.5 Windowing or Short time analysis of speech signal 24
3.6 Filter bank illustration 26
3.7 Three state left-right HMM 29
3.8 Example of a word network lattice 38
3.9 Example of a small segment of a finite state network 44
3.10 Piecewise linear warping in VTLN 48
3.11 Regression trees 51

4.1 Age distribution of speakers in the JNAS and the S-JNAS corpora . . 66
4.2 Age distribution of speakers in the training set of the SCOTUS corpus 68
4.3 Comparison of WERs on *younger adult* and *older adult* voices in the SCOTUS corpus 70
4.4 WERs (%) with increasing age on *older adult* voices in the SCOTUS corpus. 71
4.5 WERs (%) with increasing age on *older adult* voices in the SCOTUS corpus with speaker adaptation 72
4.6 WERs (%) for young and older speakers of the MATCH corpus using different language models 75
4.7 WERs (%) for young and older speakers of the MATCH corpus using different acoustic models 76

5.1 Illustration of artificial modification of fundamental frequency 83
5.2 Illustration of waveforms with artificial increase in jitter 85
5.3 Illustration of waveform with artificial increase in shimmer 87

6.1 Phoneme loop decoder 91
6.2 Phoneme correct recognition (%) on the SCOTUS corpus 93
6.3 Correct recognition (%) of most used japanese phonemes for the younger and older adults 95
6.4 Mean vowel space areas for younger and older male adults in SCOTUS corpus 96
6.5 Centroid positions of common vowels in younger and older Adults . . 98

7.1 MATCH speakers in 3D space using multi dimensional scaling on the distance matrix 108
7.2 Clustering of speakers in the MATCH corpus 109
7.3 Training gender and age dependent acoustic models 111
7.4 Unsupervised hierarchical models 113
7.5 Statistics of models chosen by test speakers in the unsupervised hierarchical models 115

8.1 Speaker Selection 123
8.2 Augmentation of adaptation data. Results on the AMI corpus 127
8.3 Augmentation of adaptation data. Results on the SCOTUS corpus for younger adult male speakers 130
8.4 Augmentation of adaptation data. Results on the SCOTUS corpus for older adult male speakers 131

List of Tables

4.1 Perplexity and OOV rate for the *younger adult* and *older adult* test sets in SCOTUS corpus 69
4.2 Comparison of the perplexities of the language model and OOV rates on MATCH corpus 74
4.3 Comparison of WERs (%) of younger and older adults in the JNAS corpus 78
4.4 WERs (%) for older adults in different age groups in the JNAS corpus 79

5.1 Fundamental frequency analysis for the phonations of vowel ‘aa’ in the SCOTUS corpus. 82
5.2 WER (%) with artificial reduction in fundamental frequency of the speech from younger adults in the SCOTUS corpus. 82
5.3 Jitter analysis for the phonations of vowel ‘aa’ in the SCOTUS corpus. 84
5.4 Jitter values computed on phonations of the vowel ‘aa’ in the original and modified waveforms 85
5.5 WER (%) with artificial increase of jitter in the speech from younger adults in the SCOTUS corpus. 86
5.6 Shimmer analysis for the phonations of vowel ‘aa’ in the SCOTUS corpus. 86
5.7 Shimmer values computed on phonations of the vowel ‘aa’ in the original and modified waveforms 87
5.8 WER (%) with artificial increase of shimmer in the speech from younger adults in the SCOTUS corpus. 88
5.9 Harmonicity analysis for the phonations of vowel ‘aa’ in the SCOTUS corpus. 88

6.1 Phonemes with largest drop in recognition rates in longitudinal study on the SCOTUS corpus 94

6.2 Vowel Space Area comparison between *younger adult* and *older adult* males in SCOTUS corpus 97
6.3 Speaking rate differences between younger and older adults on the SCOTUS corpus 99
6.4 Speaking Rate differences between younger and older adults in the JNAS corpus 100
6.5 Phoneme accuracies using phoneme loop decoder for older speakers . 102
6.6 Phoneme accuracies using phoneme loop decoder for younger speakers 102
6.7 Word correct recognition and accuracies for older speakers in the JNAS corpus with original and transition parameter modified models 103
6.8 Substitution, deletion and insertion errors for older speakers in the JNAS corpus with original and transition parameter modified models . 103
6.9 Word correct recognition and accuracies for younger speakers 103
6.10 Substitution, deletion and insertion errors for younger speakers in the JNAS corpus with original and transition parameter modified models . 103

7.1 Precision And recall for each class in age group classification task on MATCH corpus using support vector machines 107
7.2 Comparison of WERs (%) of younger and older adults in the JNAS corpus using gender dependant models 111
7.3 Comparison of WERs (%) of younger and older adults using 'Gender + Age' dependant Models 112
7.4 WERs (%) of Younger and Older odults using unsupervised hierarchical models 114
7.5 WERs (%) on older speakers in the JNAS corpus using acoustic models with modified transition parameters 117
7.6 WERs (%) on older speakers in the SCOTUS Corpus with modified transition parameters 118

8.1 Speaker identification task 122
8.2 AMI Corpus: Baseline results (WER %) 125
8.3 AMI Corpus: Results with augmented adaptation data (WER %) . . . 126
8.4 SCOTUS Corpus: Baseline results (WER %) for *younger adult* speakers 128
8.5 SCOTUS Corpus: Baseline results (WER %) for *older adult* speakers 129
8.6 SCOTUS Corpus: Results with augmented adaptation data (WER %) on *younger adult* speakers 129

8.7 SCOTUS Corpus: Results with augmented adaptation data (WER %) on *older adult* speakers 129

A.1 Comparison of WER (%) on *younger adult* and *older adult* voices in the SCOTUS corpus 139
A.2 Comparison of WER (%) using MLLR speaker adaptation on *younger adult* and *older adult* voices in the SCOTUS corpus 139
A.3 Comparison of WER (%) using vocal tract length normalisation on *younger adult* and *older adult* voices in the SCOTUS corpus 139
A.4 Comparison of WER (%) using speaker adaptive training on *younger adult* and *older adult* voices in the SCOTUS corpus 140
A.5 WER (%) with increasing age on older adult voices in the SCOTUS corpus 140
A.6 WER (%) with increasing age on older adult voices using MLLR speaker adaptation in the SCOTUS corpus 141
A.7 Comparison of WER (%) of young and older voices on MATCH corpus using different language models 141
A.8 Comparison of WER (%) of younger and older voices on MATCH corpus using different acoustic models. 142
A.9 F1 and F2 for the monophthongs in the SCOTUS corpus 142
A.10 Correct recognition (%) of phonemes on younger and older adult males in the SCOTUS corpus 143
A.11 Correct recognition (%) of phonemes on younger and older adult males in the JNAS corpus [101] 144
A.12 Speaking Rate (Frames/Phoneme) on the SCOTUS Corpus 145
A.13 Speaking rate (Frames/Phoneme) on the JNAS corpus 146

Chapter 1

Introduction

1.1 Motivation

Speech is the most natural form of communication between humans. With advances in Automatic Speech Recognition (ASR) systems, speech as a mode of communication with computing devices is finding wider acceptance in the society. Today, use of ASR can be seen in a large array of applications including interactive voice response systems such as telephone banking and ticket booking, dictation systems on personal computers, command and control in automobiles, easy dialing on mobile phones, creation of electronic medical records in health-care organisations etc.

While use of ASR systems is beneficial for everyone, it could be particularly useful for older people and especially those with mobility and visual impairments. Easy to use voice based interactive systems in health-care and home-care would make life a lot easier for them [Müller et al., 2003]. Several initiatives such as MATCH [1] and Gator Tech Smart houses [2] are focused on research and development of home care technologies and thereby to assist in independent living of the elderly people. These systems see voice as one of the important modes of interaction.

During the last century the world's ageing population has been growing at a staggering rate. According to the United Nations, in 2006, close to 500 million people in the world were aged 65 and older. Based on projections, the number will increase to 1 billion by 2030, which means one in every 8 of earth's inhabitant's will be aged 65 or above [Kevin and Philips, 2005]. This is a large segment of population and from an

[1] 'Mobilising Advanced Technologies For Care at Home' - a research project focused on technologies for home care. www.match-project.org.uk

[2] http://www.icta.ufl.edu/gt.htm

ASR research point of view it is of interest to be able to cater to their voices.

Over the years, there have been numerous studies to understand the structural changes in speech production mechanism observed with ageing. These studies have been mainly in speech pathology and speech therapy research and research associated with geratology mainly motivated by the need to understand the differences between natural changes in voice with ageing and vocal changes associated with pathological conditions. Deterioration of voice quality with ageing has been widely reported [Linville, 2001; Ramig and Ringel, 1983; Ramig et al., 2001]. Ageing also effects fine motor control capabilities and thereby the tongue movement and speaking rate. These changes impact the intelligibility of speech from older people. Cognitive abilities such as fluid intelligence, working memory span and information processing speed tend to decline as people grow old [Bäckman et al., 2001]. These cognitive factors have a large impact on the way older people interact with spoken dialogue systems [Wolters et al., 2009]. The impact of ageing on voice is also dependent on several factors specific to individuals such as their health and well being, smoking habits and their profession. These factors increase the variability and make it difficult to find a correspondence between chronological age and vocal age. All the above mentioned changes throw interesting challenges to ASR systems that need to be addressed.

ASR systems have been evolving rapidly over the last couple of decades with advances in machine learning techniques[Renals and Hain, 2010]. The problem of acoustic modeling has been studied and researched from various perspectives such as making them robust to variations in background noise, speaker characteristics, dialect and accent. From an age perspctive, there has been lot of work focused on acoustic modeling for children voices [Gerosa et al., 2009], but there has been limited work on understanding the impact of changes in acoustic characteristics associated with older voices on ASR systems. Relatively poor recognition accuracies for older voices have been reported before [Baba et al., 2004; Anderson et al., 1999; Wilpon and Jacobsen, 1996] but to the best of our knowledge there has not been an in-depth study addressing this problem. In this thesis, we address this problem and present our research work and experimental results focused on the domain of ASR for ageing voices.

1.2 Objectives

There are several components in an ASR system including the acoustic models, language models, lexicon and decoder. There is scope to adapt each of these components

in order to make the ASR systems work better for older voices. In this thesis, we address the problem from an acoustic modeling perspective.

We approach the problem from a two fold perspective. Firstly, it is of interest to analyse the changes in glottal source and articulatory characteristics of older voices and to analyse the impact of such changes on ASR recognition accuracies. Secondly, it is of interest to understand the improvements in accuracies possible with the state-of-the-art speaker adaptation techniques and to explore and propose other strategies for acoustic modeling targeted towards older voices to enhance the accuracies.

The main objectives of the thesis are outlined as follows:

1. To perform a systematic comparative study of the glottal source parameters of adult and older voices and to analyse the impact of changes in any those parameters on ASR accuracies.

2. To study articulatory changes with ageing.

3. To analyse the impact of slower speaking rate on ASR accuracies.

4. To report the baseline accuracies for older voices for a few chosen corpora.

5. To explore the possibility of speaker clustering based on gender and age group.

6. To explore the effectiveness of hierarchical models to improve the accuracies for older voices.

7. To explore the idea of improving the accuracies for a target speaker by using speech from other acoustically close speakers.

The approach to address these research objectives and the experimental results are explained in detail in the following chapters.

A couple of important factors need to be mentioned beforehand. In general, there are several disfluencies associated with very old speakers due to various pathological conditions. For the purpose of this thesis, we are mainly interested in and only investigate the speech of healthy older adults. It is also well known that chronological ageing and vocal ageing are weakly correlated. However, in this thesis we categorize speakers above 60 years of age as older adults.

1.3 Publications

Some of the ideas and results appearing in this thesis have been published in peer reviewed conference proceedings and articles during the course of this research work. Following is the list of publications and the thesis chapter in which the results of the paper are discussed.

Ravichander Vipperla, Steve Renals, and Joe Frankel. Longitudinal study of ASR performance on ageing voices. In Proceedings of Interspeech, Brisbane, 2008. (*Chapter 4*)

Ravichander Vipperla, Maria Wolters, Kallirroi Georgila, and Steve Renals. Speech input from older users in smart environments: Challenges and perspectives. In Proc. HCI International: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments, number 5615 in Lecture Notes in Computer Science. Springer, 2009. (*Chapter 4*)

Maria Wolters, Ravichander Vipperla, and Steve Renals. Age Recognition for Spoken Dialogue Systems: Do We Need It? In Proceedings of Interspeech, Brighton, 2009. (*Chapter 7*)

Ravichander Vipperla, Steve Renals, and Joe Frankel. Ageing voices: The effect of changes in voice parameters on ASR performance. EURASIP Journal on Audio, Speech and Music Processing, 2010. (*Chapter 5*)

Ravichander Vipperla, Steve Renals, and Joe Frankel. Augmentation of adaptation data. Proceedings of Interspeech, Makuhari, 2010. (*Chapter 8*)

Chapter 2

Ageing voices

In this chapter, we review the important structural and functional changes that occur in speech production when people grow old. We then review previous studies on how these changes impact voice quality and look at various measures used by researchers to analyse the quality of voice.

2.1 Human speech production mechanism

The human vocal mechanism (Figure. 2.1) consists of the lungs, the larynx (which houses the vocal cords), and the vocal tract comprised of the pharynx, the mouth and the nose.

Depending on the sound that needs to be generated, articulatory motor control mechanisms include positioning the jaw, shaping the tongue, shaping the lips, positioning the velum (to control the acoustic flow through the nasal cavity), control of the vocal cord vibrations and flow of air in and out of the lungs. As air is expelled from the lungs through the trachea, the vocal cords in the larynx are caused to vibrate by the air flow. The air flow is thus chopped into quasi periodic pulses which are modulated as they pass through the pharynx cavity, mouth cavity and nasal cavity. The combination of the shape of the vocal tract and the presence/absence of vocal cords vibrations, result in the production of various sounds [Rabiner and Juang, 1993].

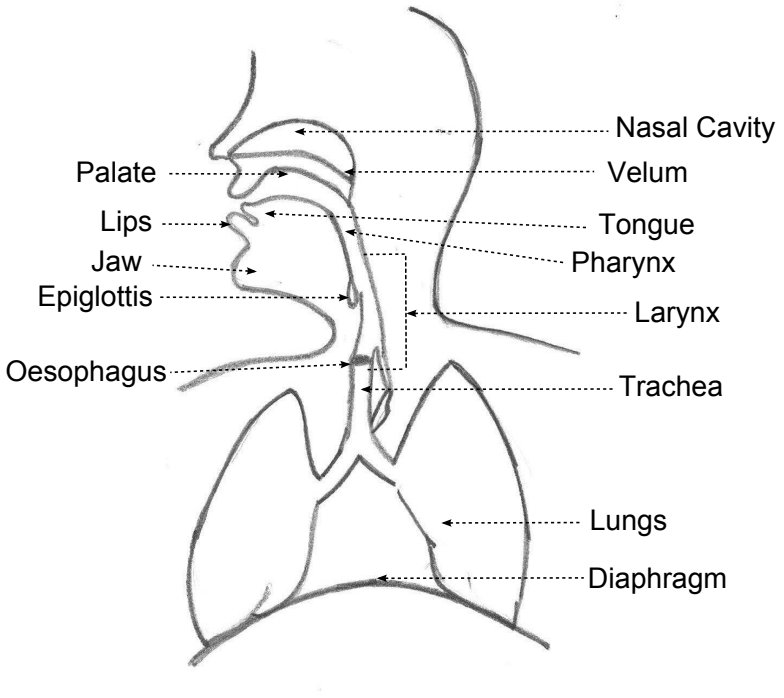


Figure 2.1: Human speech production mechanism

2.2 Changes in the speech production mechanism with ageing

Several physical and physiological changes occur in a human bodies with ageing. Typical changes include decline in vision and hearing, weakening of muscles, mobility restrictions and weakened immune system. Similar to other body parts, organs in the human speech production mechanism also undergo age related changes such as reduction in the respiratory muscle strength, restricted vocal fold adjustments during phonation and difficulty in adjustments of tongue and lip shapes [Linville, 2001]. The rate at which voices age does not however depend only on the chronological age of a person, but also on other factors such as lifestyle, physiological condition, smoking habits and profession. Even with the above mentioned factors being identical between two individuals, the extent of vocal ageing could differ between them. Described below are some of the changes seen in the voices showing signs of ageing.

2.2.1 Changes in the Respiratory system

Apart from breathing, the respiratory system plays a crucial role in producing speech. It acts as the energy source for speech production by forcing air through the vocal cords and the vocal tract resulting in various sounds.

The most significant changes seen in the respiratory system of aged people are the loss of lung elasticity, increase in the stiffness of the chest wall, and decrease in the respiratory muscle strength [Mahler et al., 1986; Rossi et al., 1996].

Lung recoil elasticity is the ease with which lungs rebound after having been stretched during inhalation. A decline in lung elasticity has been reported by Mahler et al. [1986] due to ageing. The loss of lung elastic recoil with age is found to be faster in males as compared to females [Bode et al., 1976].

Due to the alterations in the muscles of the chest wall, the thorax becomes increasingly rigid with ageing [Kahane, 1981]. This leads to a reduced movement in response to the respiratory muscle forces. Due to the degeneration of the upper and middle regions of the thoracic vertebral column, a pronounced curvature of the back is observed in some older adults. This phenomenon called Kyphosis, alters the shape of the thorax and may effect the amount of air that can be inhaled and exhaled.

Several research studies have reported weakening of respiratory muscles during old age [Black and Hyatt, 1969; Kahane, 1981]. This leads to reduced respiratory forces during inhalation and exhalation. A decline in maximal respiratory pressure progressively beyond the age of 65 has been reported by Enright et al. [1994]. The decline is more prominent in males compared to females. A loss in diaphragm strength leading to an average reduction of 25% of maximum transdiaphragmatic pressure in elderly group as compared to younger subjects has also been reported [Tolep et al., 1995].

While the total lung volume remains unaltered in the older people, the forced expiratory volume and the lung pressure are decreased. This leads to a decline in the amount of air that can be moved in and out of the lungs and the efficiency with which it can be moved [Linville, 2004; Ramig et al., 2001]. The rate of this decline accelerates with advancing age [Mahler et al., 1986]. Also the amount of air left after exhalation known as ‘Residual volume’ has been found to increase by about 40% from the age of 20 to the age of 70 [Lynne-Davies, 1977].

2.2.2 Changes in the Larynx

The parts of the larynx that form the vocal apparatus are the laryngeal cartilages (to which the vocal folds are attached), the vocal folds that play a key role in phonation, and the intrinsic muscles that regulate the vocal cord tension and the vocal fold opening [Pretterklieber, 2003]. Several anatomical changes are seen in these organs with ageing.

Among the several cartilages in the larynx, the thyroid, cricoid and arytenoid cartilages are the most significant from the speech production point of view. The thyroid and cricoid cartilages form the skeleton of the larynx. A pair of arytenoid cartilages are located on the upper edge of the cricoid cartilage. The vocal cords are attached posteriorly to the arytenoid cartilages and anteriorly to the thyroid cartilages. The cricoarytenoid joints allow the arytenoid and thus the vocal apparatus to move laterally or medially. The arytenoids can also glide on the surface of the cricoid and move closer or recede away from each other. The most significant change in the cartilages observed as an individual moves from adulthood to old age is the toughening of the soft tissue into bone like structure (ossification). This phenomenon is observed in both males and females. It occurs at an earlier age and is more prominent in males as compared to females. Each of the cartilages has its own pattern of ossification. Arytenoid cartilage ossifies only partially sparing the vocal process. Significant age-related changes have been reported in the cricoarytenoid joint [Paulsen and Tillmann, 1998; Dedivitis et al., 2001]. Changes include thinning of the joint surface, reduced collagen fibers in the cartilage matrix and surface irregularities. These changes are again more prominent in males compared to females and hamper overall positional or postural movements of the arytenoid cartilages. This leads to reduction in the degree and extent of vocal ligament closure and makes it difficult for vocal fold adjustments during phonation. The result of this is impaired vocal quality and reduced vocal intensity due to air leakage through incomplete vocal fold closure.

The vocal folds have a complex layered structure. They are comprised of five discrete histological layers: the Epithelium, three layers jointly called Lamina Propria and the Thyroarytenoid muscle. The thin layer of Epithelium forms the protective covering for the vocal folds. The epithelial cells are bound together firmly and form a smooth lining reducing the friction to the air flow. The superficial layer of Lamina Propria is a thin layer made of elastin fibres. This layer can be stretched in several directions. The intermediate layer which is formed of elastin and collagen fibres is

more densely packed and can only be stretched in anterior-posterior direction. The deep layer is formed on collagen fibres and is least stretchable. This layer protects the vocal cords from over extension. The Thyroarytenoid muscle lies below the Lamina Propria. They are mainly concerned with pulling together the thyroid and arytenoid cartilage, thus relaxing the vocal folds.

Several changes in the structure with ageing alter the biomechanical properties of the vocal folds [Linville, 2001]. Glandular changes in the laryngeal mucosa (the mucous lining of larynx) [Linville, 2004] cause drying of the epithelial tissue, increasing the stiffness of vocal cord cover. This increase in cover stiffness leads to instability of vocal fold vibration. Some investigations [Hirano et al., 1989] have reported thickening of laryngeal epithelium progressively with age. Tissues age at varying rates and to varying extents [Kahane and Hammons, 1987] and substantial structural changes need to occur before observing noticeable changes in voice.

In the Lamina Propria, several age related changes have been documented in all the three layers. The thickness of the superficial layers alters [Hirano et al., 1989] and atrophy and degeneration of the elastic fibres in the layer has been observed [Sato and Hirano, 1997]. Changes seen in the intermediate layer include thinning of the layer, decrease in the density of the fibres, atrophy of the fibres and changes in the contour of the layer [Linville, 2001]. The fibrous protein loses elasticity and the layer stiffens. The deep layer thickens with an increase in the collagen fibres. Such morphological changes in the fibres of the vocal folds contribute partially to the ageing of the voices.

The thyroarytenoid muscle also displays atrophy with ageing. Changes in muscle fibres have been reported [Sato and Tauchi, 1982]. A decrease in thyroarytenoid muscle activity has been reported [Baker et al., 1998] in older speakers than young speakers. This affects the fine control of the position of the arytenoid joint and thereby the fine control of pitch of the voice.

Intrinsic laryngeal muscles are responsible for control of the vocal cords. The tension in the vocal cords is regulated by the cricothyroid muscle. The opening (abduction) of the vocal fold opening (called Rima Glottidis) is controlled by the posterior cricoarytenoid muscle and the closing (adduction) is controlled by the lateral cricoarytenoid and thyroarytenoid muscles. Regressive changes and atrophy have been reported in all these muscles with ageing [Rodeño et al., 1993; Bach et al., 1941]. The changes include accumulation of fats, degeneration of muscle fibers and unusual variations in the cross sectional areas [Linville, 2001]. As a result, precise control of the vocal cord tension and complete abduction/adduction is affected.

2.2.3 Changes in the vocal tract

The human vocal tract consists of all the organs above the vocal folds that are involved in speech production. It is comprised of the pharynx (throat), the oral cavity, the nasal cavity, soft palate (velum) and the articulators viz., the tongue and the lips. The human speech production mechanism can be viewed as a source-filter model. The lungs in conjunction with vocal cords act as the source and expel air into the vocal tract. Depending on the presence or absence of the vocal cord vibrations, the source is either voiced or unvoiced. This quasi periodic air then resonates in the pharynx, oral and nasal cavities to generate a rich timbre. The vocal tract thus acts as the filter.

The vocal tract can be broadly thought to be comprised of three resonating cavities, the pharynx, and the oral and nasal cavities. The pharynx is involved in the production of all speech sounds. The pharynx can change shape to a limited extent and thus alter the resonance patterns. The pharynx can be constricted, and it can be raised or lowered. The position of the velum also alters the shape of the pharyngeal cavity. The velum controls the flow of air into the nasal cavity. During the production of nasal sounds such as /m/ and /n/, the velum is moved forward to open the air passage through the nasal cavity. The oral cavity is the most flexible among the three cavities in varying the shape. The resonating property of the oral cavity depends on the position of the temporomandibular joint, the shape of the tongue and the lips and the position of the velum.

Thinning of pharyngeal epithelium and degeneration of the pharyngeal muscles has been reported with ageing [Linville, 2001]. However these changes in the pharynx are not found to be extensive.

The temporomandibular joint (TMJ) is the joint at which the jaw is hinged to the skull. It is used in controlling the position of the jaw and hence influences the oral resonance during speech production. Jaw movement has a significant role to play in articulation of certain phonemes as well as in the co-articulation of adjacent phonemes. With ageing, degenerative changes are observed in the TMJ [Weinstein, 2000]. Displacement of the TMJ disk is commonly observed leading to a lowering of the articulating surface. Xue and Hao [2003] have reported increase in vocal tract dimensions in older speakers. The vocal tract volume of older speakers in particular is significantly higher compared to the younger speakers. This could lead to changes in the resonance patterns in older voices.

The tongue plays a major role in speech production. It is very flexible and can be

moved up, down, forward and backward. By adjusting the shape of the tongue and the position of the tongue tip, the oral cavity's shape is modified affecting the resonance patterns and hence the sound produced. Significant changes have been reported in the tongue with ageing [Rother et al., 2002]. Decrease in the thickness of epithelium and glandular atrophy have been reported in people over 50 years of age [Nakayama, 1991]. However the most significant change in the tongue that affects the speech production is the atrophy of the tongue muscles. From ultrasound observations, decline in the tongue motor skills in the elderly in comparison to young adults were reported by Koshino et al. [1997]. A decline in tongue strength has also been reported in older individuals [Crow and Ship, 1996]. These changes in the tongue could affect the articulatory patterns.

Other changes observed in the mouth with ageing include loss of oral mucosa (the mucous membrane that covers all the structures inside the oral cavity other than the teeth), decline in the salivary function leading to oral dryness and degeneration and loss of tooth. These changes could also have a small impact on speech production.

2.2.4 Neuromuscular control

Age related changes also take place in the peripheral and central nervous system that have implications for speech production. One of the changes in the peripheral neural system is the decline of motor neurons. This loss in the motor units has been implicated as the primary mechanism for muscle atrophy and loss of contractile strength in the muscles [Doherty et al., 1993]. An average loss of 25% neurons has been reported from the second to the tenth decade of life. However this loss of motor units is partially compensated by increase in the size of the motor units along with the slowing of contractile speed. This affects various muscles involved in the speech production and is a possible cause of the slower speaking rate observed in older speakers.

Age related memory impairment is commonly observed in elderly people [Hedden and Gabrieli, 2004]. In particular reduction in working memory and the associated difficulty in refreshing recently processed information have implications on speech production behavior and interaction styles.

2.3 Acoustic effects of ageing

Several studies have been made to understand the effect of ageing on various acoustic parameters of speech. These studies have been mainly in the field of speech pathology to differentiate normal voice changes due to ageing from pathological vocal conditions affecting elderly patients. Most of these studies [Ramig and Ringel, 1983; Ramig et al., 2001; Linville, 2000; Edward, 1959] have indicated that speakers experience certain changes, mainly deterioration, of vocal acoustic output as they age.

To analyse the voice quality, different parameters of speech signal have been proposed and widely used. This section provides a brief description of the parameters that have been used in this thesis. Some of these parameters such as the fundamental frequency, jitter and shimmer relate to the characteristics of the glottis and hence can be treated as source related parameters. Other parameters, such as formant frequencies and speaking rate relate to the shape and movement of the vocal tract and are thereby treated as filter related parameters. Although these parameters have been primarily used to differentiate between healthy voices and those suffering from pathological conditions, they have also been used to study the change in voice quality with ageing. These parameters are typically measured on sustained phonations of few seconds in duration recorded in noise free sound booths.

2.3.1 Average fundamental frequency

Among the several parameters affected by ageing, the average fundamental frequency (F_0) has been one of the most extensively studied parameters. Although there is no general agreement on the trend, it appears [Schötz and Müller, 2007; Linville, 2000] that in females, the fundamental frequency remains fairly constant until menopause, and later decreases. A drop of approximately 10-15 Hz is observed. This is attributed to the thickening of laryngeal mucosa. while in males F_0 decreases until a certain age around 60 years and increases after that significantly. However the experiments in [Xue and Deliyski, 2001; Endres et al., 1971] indicate that F_0 reduces significantly for both the males and females. A decrease of 40-60 Hz in F_0 has been reported for both males and females.

2.3.2 Fundamental frequency variation and Amplitude variation

Older voices are generally associated with tremor and increased hoarseness. These characteristics are related to F_0 and amplitude instability. Measures of standard deviation of the fundamental frequency and its amplitude, indicate gross stability of voice over time. These measures tend to increase with age for both males and females [Linville, 2000]. The F_0 standard deviation more than doubles between young adulthood and old age for men while an increase of over 70% has been observed in older women's voices. These observations are also confirmed experimentally by Xue and Deliyski [2001]; Bruckl and Sendlmeier [2003].

2.3.3 Jitter

Jitter is the cycle to cycle variation of the pitch period, i.e., the average of the absolute distance between consecutive periods. It is measured in μsec.

$$Jitter\ absolute \qquad \frac{1}{N-1}\sum_{i\ \ 1}^{N-1}|T_i - T_{i\ \ 1}| \tag{2.1}$$

where T_i is the extracted F_0 period length and N is the number of extracted F_0 pitch periods [Boersma, 2001].

A relative measure for frequency perturbations known as 'Jitter Local' is often used. It is the ratio of pitch period variation from cycle to cycle to the average pitch period. It is expressed as a percentage.

$$Jitter\ Local \qquad \frac{\frac{1}{N-1}\sum_{i\ \ 1}^{N-1}|T_i - T_{i\ \ 1}|}{\frac{1}{N}\sum_{i\ \ 1}^{N} T_i} \tag{2.2}$$

The other measures of jitter that are averaged over larger number of pitch periods are as follows:

Relative Average Perturbations (Jitter RAP): The average absolute difference between a period and the average of it and its two neighbours, divided by the average period.

Five point Period Perturbation Quotient (Jitter PPQ5): The average absolute difference between a period and the average of it and its four closest neighbours, divided by the average.

Difference of differences between periods (Jitter DDP): The average absolute difference between consecutive differences between consecutive periods, divided by the average period.

Increased jitter with age has been observed in both males and females. But it has been suggested by Ramig and Ringel [1983]; Linville [2001] that amplitude perturbation measures may be better discriminators of age than cycle-to-cycle variations. Jitter is caused by the instability of the vocal folds. With ageing, due to physiological changes and deterioration in health, the vocal folds may weaken causing jitter, but if the older person is in a healthy condition, the difference in Jitter from a young adult does not differ too much [Linville, 2001]. Though Jitter may not be a clear indicator of chronological age, it does provide some acoustic cues to indicate ageing.

2.3.4 Shimmer

Shimmer is the variability of the peak-to-peak amplitude in decibels. It is the ratio of amplitudes of consecutive periods. It is expressed as

$$\textit{Shimmer dB} \qquad \frac{1}{N-1}\sum_{i\ 1}^{N-1} |20log\ \frac{A_{i\ \ 1}}{A_i}\ | \tag{2.3}$$

where A_i is the peak-to-peak amplitude in the period and N is the number of extracted fundamental frequency periods.

Relative shimmer (Shimmer Local) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage.

$$\textit{Shimmer Local} \qquad \frac{\frac{1}{N-1}\sum_{i\ 1}^{N-1}|A_i - A_{i\ \ 1}|}{\frac{1}{N}\sum_{i\ 1}^{N} A_i} \tag{2.4}$$

Similar to the Jitter measurements, Shimmer is also measured by averaging over larger number of periods.

Three point Amplitude Perturbation quotient (Shimmer APQ3): The average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

Five point Amplitude Perturbation Quotient (Shimmer APQ5): The average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.

Difference of difference between amplitudes (Shimmer DDA): The average absolute difference between consecutive differences between the amplitudes of consecutive periods.

Shimmer has been found to have a strong correlation with age. Xue and Deliyski [2001] found the mean shimmer in older voices to be 0.48 dB while the value was 0.19 dB for younger speakers. Amplitude perturbations have also been reported to increase with age by Ramig and Ringel [1983]; Linville [2000]; Bruckl and Sendlmeier [2003]; Bruckl [2007]. Shimmer levels increase with age independent of health and fitness variables, and hence serves as a good indicator of ageing voice.

2.3.5 Breathiness

Another voice quality associated with ageing is increased breathiness. In general women are judged to be breathier than men. Breathiness is thought to be due to the incomplete glottal closure during closed phase of the phonatory cycle. The nearly sinusoidal shape of the breathy glottal waveforms is responsible for increase in the relative amplitude of the first harmonic. It has also been observed in [Kilch, 1982] that breathy signals tend to have more high-frequency energy than normally phonated signal. Another property of breathy signals are that they are less periodic, especially in mid and high frequencies where aspiration noise is large [Hillenbrand et al., 1994].

Harmonic to Noise Ratio (HNR) measures the signal to noise ratio in a periodic waveform and acts as a good indicator of voice quality. It is computed as the logarithm of the ratio of the energy of the signal in the periodic part to the noise. It is measured in decibels [Boersma, 1993]. HNR was found to be a sensitive index of vocal function [Ferrand, 2002] and a significant lowering of the HNR values were reported in older voices.

A measure that correlates well with breathiness in voice is Cepstral Peak Prominence (CPP) proposed by Hillenbrand and Houde [1996]. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of a signal. When the log amplitude of the spectrum contains regularly spaced harmonics, the Fourier analysis of the spectrum then captures the periodicity in the spectrum and will show a peak at a quefrency corresponding to the spacing between the harmonics. The cepstral peak reflects both the level of harmonic structure in the signal and the overall amplitude of the signal. To normalise for overall amplitude, a linear regression line is calculated relating quefrency to cepstral magnitude. The CPP measure is the difference in amplitude (in dB) between the cepstral peak and the value of the regression line at the cepstral peak (Figure 2.2). CPP is computed on frames of 10 ms and averaged over all the frames in an utterance. CPP values for breathy voices are lower than those for normal voice since the cepstral

peak is expected to be smaller in breathy voices due to loss of periodic structure in higher frequencies of the spectrum.

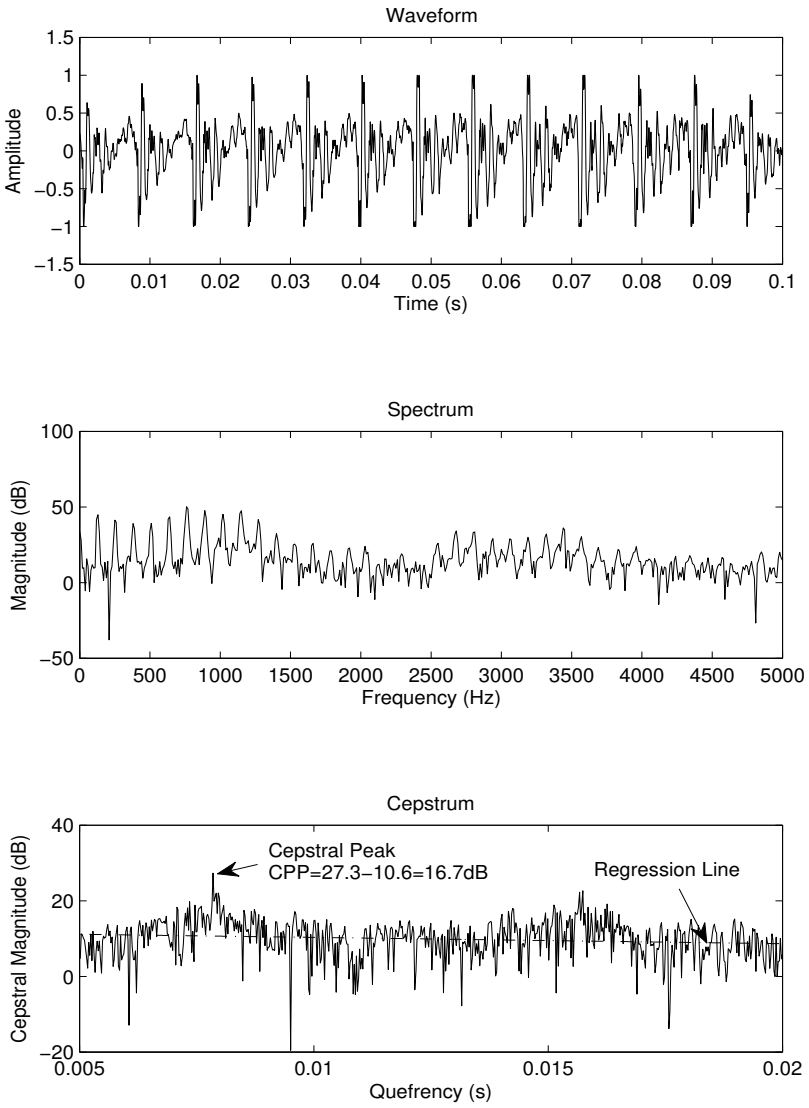


Figure 2.2: Cepstral peak prominence

A smoothed version of CPP called CPPS is computed similarly with some additional smoothing. For CPPS, a frame size of 2ms is used instead of 10ms and two levels of smoothing are applied. First the cepstrum is averaged across time by replacing an unsmoothed cepstrum at a time frame with the average of itself and the adjacent cepstral frames. A second level of smoothing is then applied by a running average of the cepstral magnitude across quefrency for each cepstral frame.

2.3.6 Sound pressure level

Increase in sound pressure level in conversational speech has been observed in males over 70 years of age, while no noticeable changes have been observed in females [Linville, 2001]. Typically with ageing there is a decline in hearing capabilities. People usually adjust the sound level based on feedback from internal coupling of sound production and sound perception. Increase in voice sound pressure level with ageing is believed to be a compensation mechanism to overcome the hearing loss. However increase in speech intensity has been observed even for older speakers with no hearing loss.

Experiments comparing long term spectral amplitudes of older adults with those of younger adults [Linville, 2002] show a significant increase in amplitudes at 160 Hz and frequencies above 6000 Hz. These findings were associated with increased breathiness in the voices.

2.3.7 Speech rate

Older speech is characterised by a lower speech rate. The lowering of the speech rate is due to the degeneration of the muscles and reduced efficiency of the peripheral motor system. It has also been suggested that older speakers deliberately slow down their speech in response to the restrictions in the functionality of the articulators in order to be more intelligible.

The speech rate is related to segment duration, number of segments per unit time and duration and frequencies of pauses. The number of speech units (syllables, phonemes, sub-phonemes etc) per second generally decrease with age [Schötz and Müller, 2007]. Several studies have reported a decrease of 20-25% in speech rate in older speakers reading and speaking rates. Increase in vowel and consonant durations and an increase in pause durations and frequencies [Bruckl and Sendlmeier, 2003; Linville, 2001; Schötz and Müller, 2007] have been reported. Speech rate reduction with age has been found to be more prominent in men than in women.

Perceptual tests on age recognition [Harnsberger et al., 2008] suggests that speaking rate is used as a strong cue in distinguishing older speakers from younger speakers.

Chapter 3

Automatic Speech Recognition

Automatic speech recognition systems attempt to transcribe a speech signal to a string of words. Given the intrinsic variability in speech due to differences in environment, speaker accent, gender, age, and emotions, this is not a straightforward task. Decades of research have not yet been able to make machine based speech recognition comparable to human performance. However, state-of-the-art ASR systems achieve good recognition accuracies on constrained tasks including simple isolated word recognition tasks on few hundred words and continuous speech recognition on larger vocabularies of the order of 50000 words.

Several models, theories and algorithms from Mathematics, Computer science and Linguistics form the basis on which the current ASR systems are built. Digital signal processing techniques, probabilistic models, machine learning techniques, finite state automata, formal logic and grammar representations, language, linguistic and phonetic knowledge find their way into the design of various components of the speech recognition systems.

3.1 ASR architecture

In order to be able to recognise the input speech, it is first parametrised into a sequence of equally spaced discrete feature vectors O as shown in Figure 3.1. Given the observed feature vectors, the basic decision rule used to hypothesize the spoken word sequence '$\widehat{w}$' is given by

$$\widehat{w} \quad \arg\max_{w} \quad P\ w|O \tag{3.1}$$

where,

$O = \{o_1, o_2, ..., o_T\}$ is the sequence of speech feature vectors (observations).

$P(w|O)$ is the probability of a word sequence w given the observation sequence O.

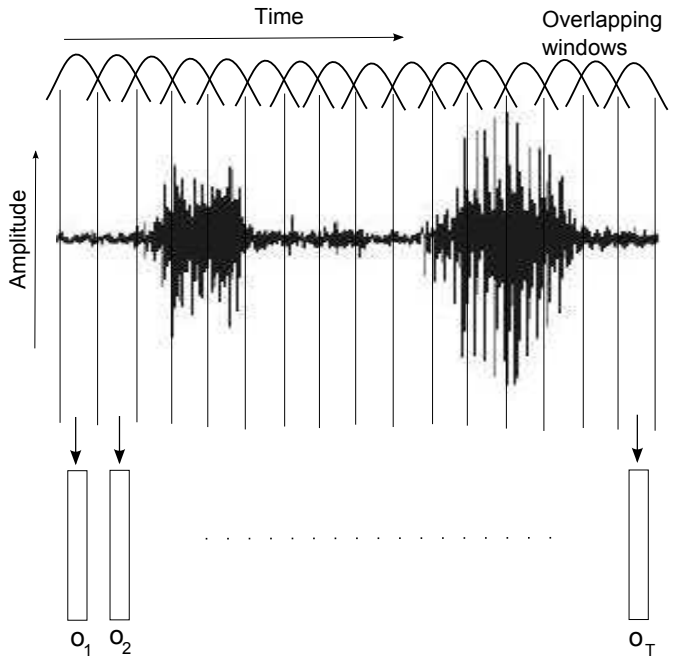


Figure 3.1: Parametric representation of speech

To keep the computation tractable, equation 3.1 is rewritten using Bayes' rule as follows:

$$\hat{w} = \arg\max_{w} P(w|O) = \arg\max_{w} \left\{ \frac{P(w)\,P(O|w)}{P(O)} \right\} \tag{3.2}$$

where,

$P(w)$ is the a-priori probability of the word sequence w.

$P(O|w)$ is the probability of observing parameter vectors O given the word sequence w

State-of-the-art ASR systems (Figure. 3.2) are built around this mathematical formulation. The probability of a word sequence $P(w)$ is computed from language models. Acoustic models are used in the computation of the likelihood $P(O|w)$. The lexicon acts as a map between the words in the language model and the sub word units that comprise the acoustic models. The feature extraction module assumes the task of converting the speech signal into discrete parameter vectors.

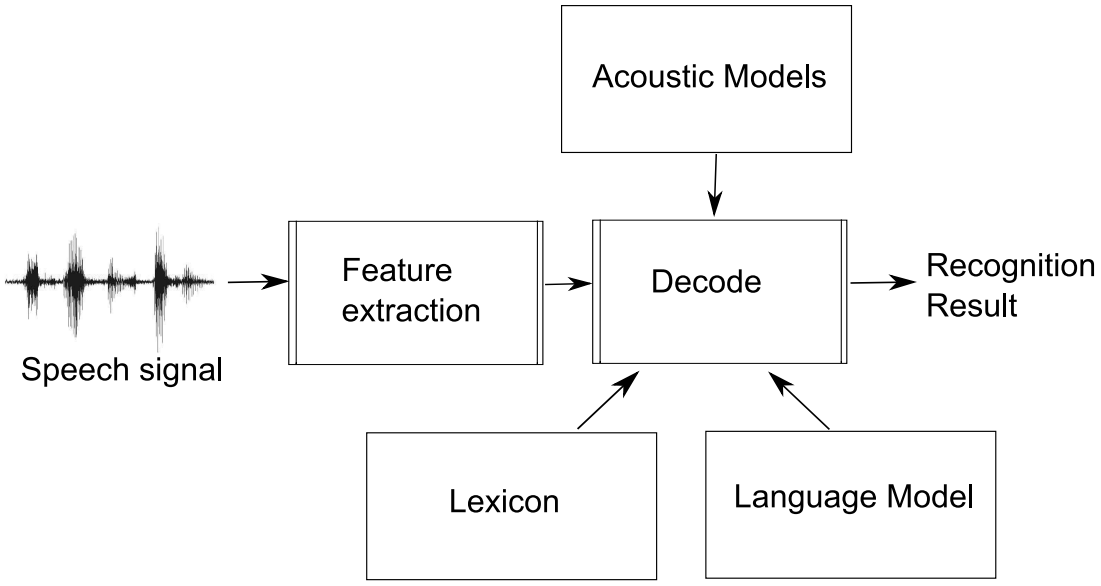


Figure 3.2: Automatic speech recognition system

3.1.1 Feature extraction

A speech signal $s\,t$ captured from a microphone is typically processed through an Analogue to Digital (A/D) converter to get a sequence of samples $s\,n$ representing the original speech. The goal of the feature extraction module is to extract meaningful features from this signal such that the features provide

1. a compact representation of the original signal.
2. good discrimination capacity between different speech sounds.
3. robustness against noise.
4. minimal variations due to speaker characteristics

Several prominent features of the speech signal, speech production and speech perception are taken into consideration in the design of feature extraction techniques for ASR.

1. The speech production mechanism can be represented as a source filter model, where the air pushed from the lungs with or without vocal fold vibrations can be treated as the source with the vocal tract acting as a filter shaping this source signal. Figure 3.3 illustrates the speech signal, its spectrum and the frequency response of the vocal tract. The vocal tract's shape leads to resonance at certain

frequencies known as the formants. By varying the shape of the vocal tract, the frequencies of the formants can be controlled thereby generating various sounds. It is hence of interest to model the underlying vocal tract shape to represent the speech signal.

2. The spectral slope of the speech signal is found to be negative. Lower frequencies have higher amplitude compared to higher frequencies.

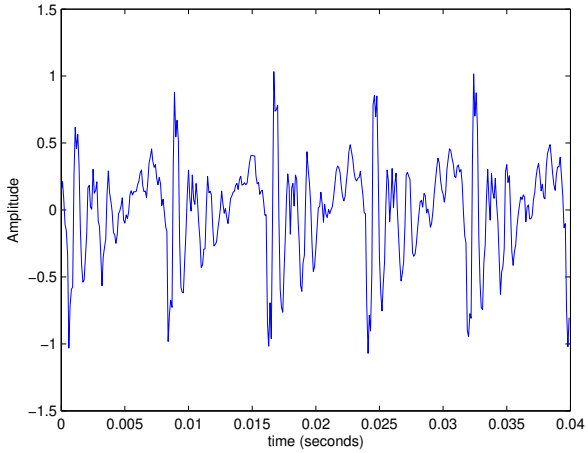
3. Though speech is a non-stationary signal, it is a reasonable to assume stationarity over short durations of 20-30 msecs.

4. The human ear frequency resolution is non-linear with respect to frequency. The relation between actual frequency and the perceived frequency is logarithmic in nature as reported in [Stevens et al., 1937].

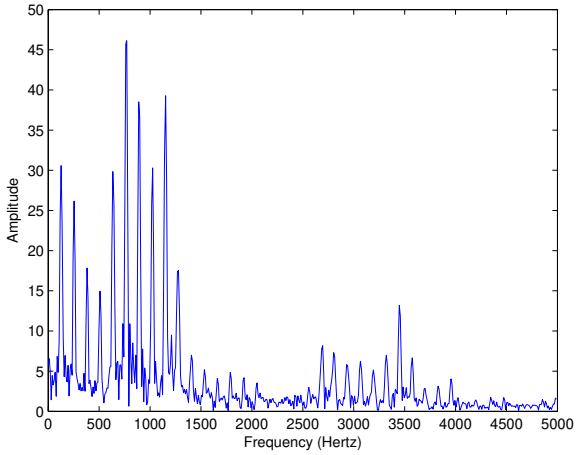
5. The human ear response is also non-linear in sensitivity to amplitude/sound pressure at different frequencies. [Robinson and Dadson, 1956; ISO-226:, 2003]

Several feature extraction techniques inspired by auditory and perceptual models combined with machine learning techniques to reduce the dimensionality and to reduce the correlation across dimensions in the feature space have been proposed to date. Some of the earliest feature extraction techniques were based on linear prediction analysis where the vocal tract's spectral response is modeled as an all-pole filter and the filter coefficients formed the feature vector. Approaches based on analysing the speech signals in the frequency domain proved to be more effective. The Cepstrum (spectrum of log spectrum) was proposed [Bogert et al., 1963] as an effective tool for homomorphic speech signal processing. State-of-the-art feature extraction techniques viz., Mel Frequency Cepstral coefficients (MFCC) [Mermelstein, 1976; Davis and Mermelstein, 1980] and Perceptual Linear Prediction coefficients (PLP) [Hermansky, 1990] use the cepstral analysis. Instead of using a fixed frame length for speech analysis, other approaches using variable frame length and rate [Bridle and Brown, 1982] and wavelets have been tried with limited success.

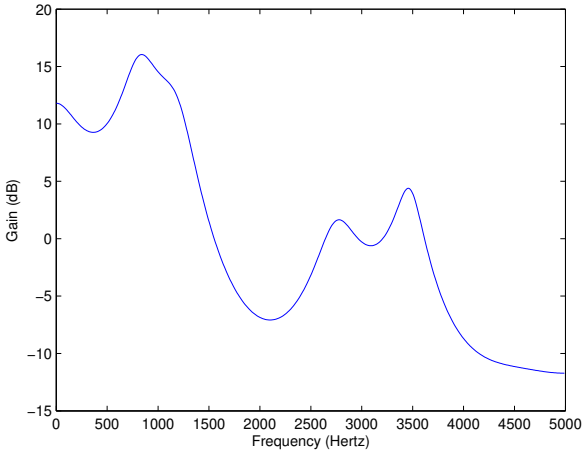
MFCCs and PLPs, as implemented in HTK [Young et al., 2006] have been used in this thesis. Both these methods are based on mel filter bank analysis. PLPs incorporate properties of human ear perception into the feature extraction process. Figure 3.4 shows a block diagram representing the various steps involved in the MFCC and PLP generation. The motivation and the process involved in each step is described below.



(a) Speech waveform



(b) Spectrum



(c) Vocal tract response

Figure 3.3: Speech signal and vocal tract response. a) shows a voiced segment of a speech waveform b) Discrete Fourier transform of the waveform c) Frequency response of the linear predictor filter modeling the speech waveform

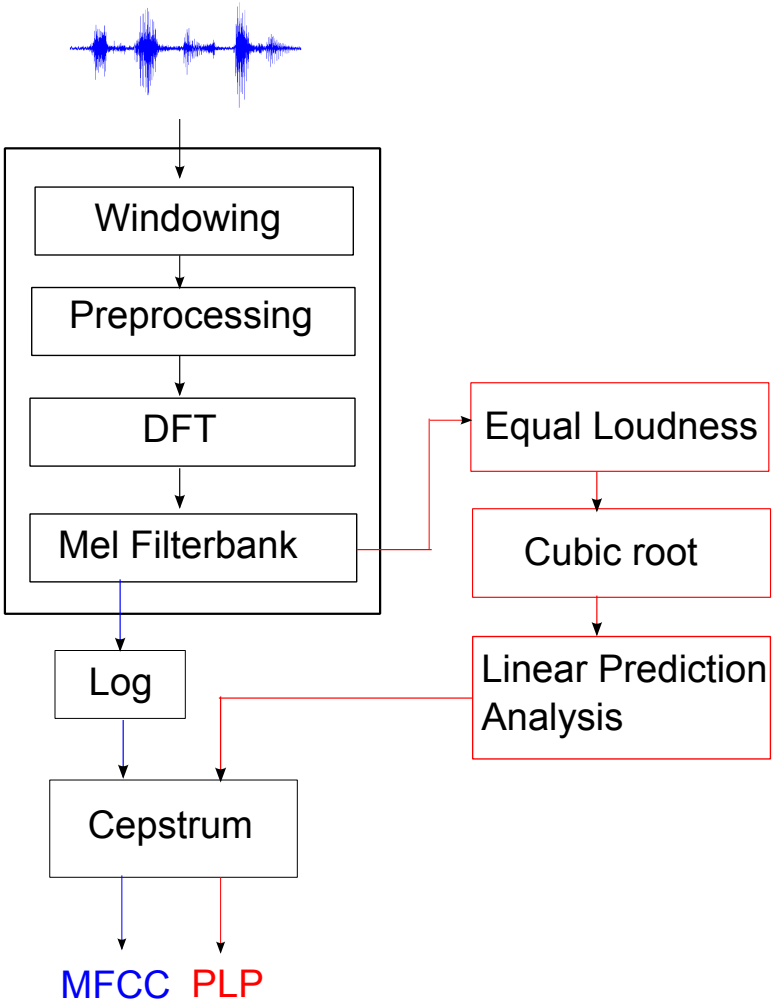


Figure 3.4: MFCC and PLP feature extraction

3.1.1.1 Windowing

In order to do frequency analysis of a signal, it needs to be stationary. As discussed above, a speech signal is assumed to be stationary over short intervals of time. Hence a sliding window approach is used, with overlapping adjacent frames as shown in figure 3.5 . Typically a window size of 25 msec and a frame shift of 10 msec are used.

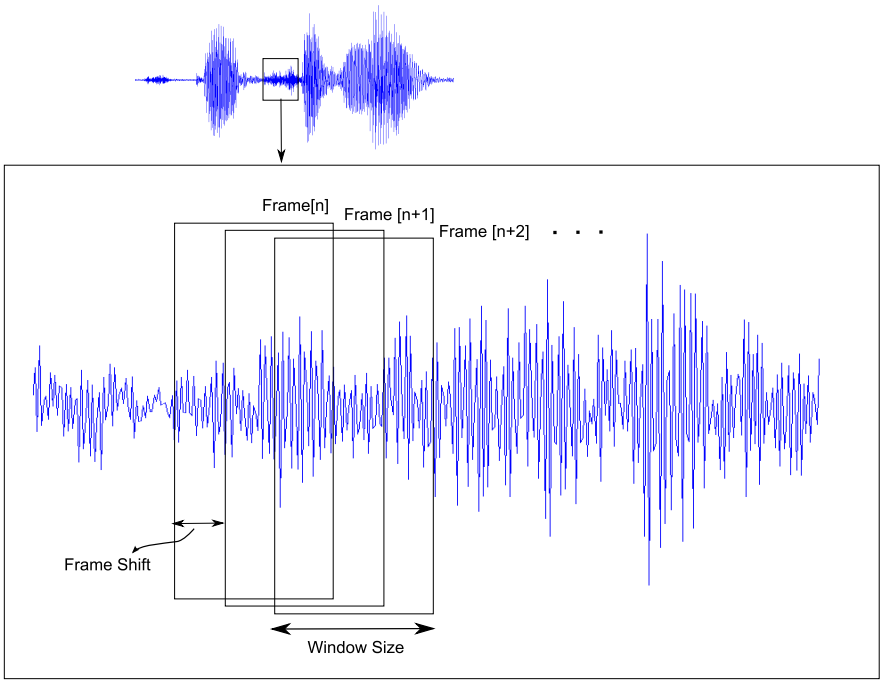


Figure 3.5: Windowing or Short time analysis of speech signal

3.1.1.2 Pre-emphasis

The speech signal is typically preprocessed before the actual feature extraction. Any DC offset introduced by the A/D converter is first removed. The signal is then pre-emphasised to boost the signal signal strength in the higher formants, using a high pass filter:

$$H(z) = 1 - kz^{-1} \qquad (3.3)$$
$$0 \leq k < 1$$

The pre-emphasis coefficient k is typically chosen close to 1 and a value of 0.97 has been used for all the experiments in this thesis.

3.1.1.3 Conversion to frequency domain

Using rectangular windowing is equivalent to convolving the speech signal with a sinc function which introduces overtones of the signal at higher frequencies. To avoid this, a smoothing window function is applied [Harris, 1978]. A raised cosine (hamming window) is often used in speech processing due to its capacity to maximally suppress the overtone frequencies.

$$\hat{s}\ n \quad \left(0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right) s\ n \tag{3.4}$$

Each frame is then converted to frequency domain using the Discrete Fourier Transform.

3.1.1.4 Filterbank analysis

To replicate the human ear resolution of the frequencies which is logarithmic in nature, the frequency domain signal is transformed from a linear scale to a logarithmic mel scale [Stevens et al., 1937].

$$f_{mel} \quad 1127\log_{\ e}\left(1 \quad \frac{f}{700}\right) \tag{3.5}$$

This is achieved in practice using triangular overlapping windows as shown in Figure 3.6. The energy in each frequency bin (m_j) is accumulated by weighting the spectral amplitude of the original signal by the value of the corresponding triangular filter at that frequency. This gives a lower dimensional feature vector (equal to the number of frequency bins). The width of the windows increase as the frequency increases in correlation to the mel scale. This approach is particularly efficient, since it provides a larger bin for higher frequencies where the energy is low.

3.1.1.5 Cepstral analysis

The goal of cepstral analysis is two fold:

1. The mel filter bank coefficients are not decorrelated due to the overlapping frequency bins. We would however like the feature vectors to be independent across dimensions.

2. The mel filter bank coefficients represent the frequency components of the speech signal which is mathematically a convolution of the glottal source signal and the

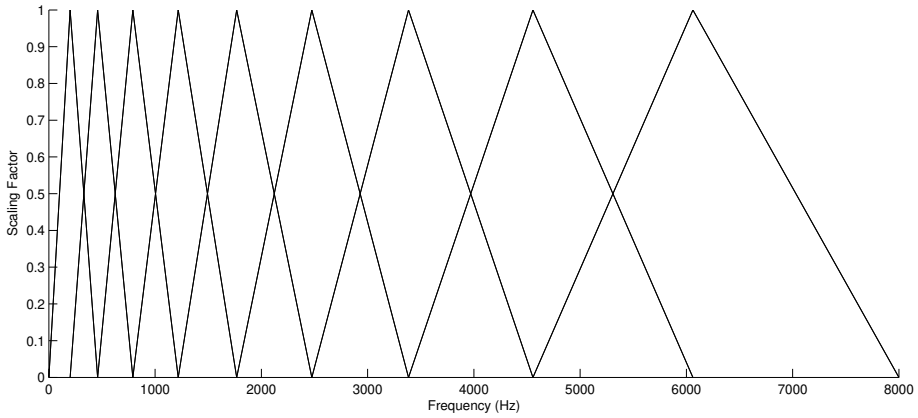


Figure 3.6: Filter bank illustration (using 9 bins for a signal sampled at 16KHz)

vocal tract channel response. It is desirable to somehow suppress the glottal source characteristics as much as possible and only capture the vocal tract formants.

Conversion to Cepstral domain involves a discrete cosine transform on the log of the mel filter bank coeffients (m_j).

$$c_i \quad \sqrt{\frac{2}{N}} \sum_{j\ 1}^{N} \log m_j \cos\left(\frac{\pi}{N}\ j - 0.5\ i\right) \tag{3.6}$$

This can be viewed as projecting the signal onto an orthogonal basis. Hence the resulting coefficients are decorrelated. Convolution in the time domain is equivalent to multiplication in frequency domain, and this multiplication becomes a simple addition in log frequency. Hence by subtracting the cepstral mean in the log cepstral domain, the glottal source characteristics can be suppressed as well.

3.1.1.6 Equal loudness

Inspired by the studies on human auditory system [Robinson and Dadson, 1956], the signal is pre-emphasized according to the equal loudness curves of the human ear response. Usually a piecewise linear approximation of the equal loudness curves is used in this process. This scaling is applied to the mel filterbank coefficients as implemented in HTK. In the original proposal of PLP [Hermansky, 1990], the frequencies are warped using a bark scale instead of mel scale.

3.1.1.7 Cube root amplitude scaling

The amplitude of the speech signal and perceived loudness are not linearly related. An empirical relationship between the two was proposed in [Stevens, 1957] called the psychophysical power law or Stevens law. As per this law, a cube root compression is applied to the energy in all the frequency components in the signal.

As a result of this step, the variations in the spectral amplitudes reduce and this gives an additional advantage in the following linear prediction step since the signal can be modeled by a lower order filter.

3.1.1.8 Linear prediction analysis

An all pole filter model is used to compactly represent the vocal tract frequency response.

$$H\ Z \qquad \frac{1}{1 \quad \sum_{i}^{p}{}_{1} z^{-i}} \tag{3.7}$$

An all pole filter is a good model to represent speech since the high energy formants can be captured by the location of the poles on the frequency axis. In practice, a filter of the order of 10 to 15 is sufficient to efficiently model the speech signal. In linear prediction analysis we attempt to compute the coefficients a_i of the filter such that the mean square error between the original speech signal $s\ n$ and the predicted signal $\hat{s}\ n$ is minimised over all the speech samples in the current frame of analysis. Using an autocorrelation method, this optimisation problem can be posed as the problem of solving M equations with M unknowns [Makhoul, 1975] which has $O\ n^3$ complexity. A more efficient method known as Levinson-Durbin method [Levinson, 1947] that exploits the Toeplitz structure of the autocorrelation matrix is commonly used due to its better $O\ n^2$ computational complexity.

3.1.1.9 Energy and differential coefficients

After the MFCC/PLP feature extraction, the log of the energy of the signal in the current frame is usually appended to the features. Throughout the feature extraction, all the processing has been done under the assumption that each frame/window of speech is independent of others. However, this is not true. There is a high degree of correlation between speech frames close to each other. To capture this dynamics in speech signal, additional first and second order differential coefficients computed using the static features from the set of adjacent frames are also usually appended to the static features.

3.1.2 Acoustic models

Acoustic models aim to compactly represent the speech sounds as mathematical models. In practice, models are usually built at the phoneme level. These phoneme models can be concatenated to model a word or an utterance.

In fact the speech signal cannot be assumed to be stationary even within a phoneme unit since each phoneme realisation can be approximated as beginning of the phoneme, steady state and the end of the phoneme. Hence under the assumption that the signal is stationary in each of these phases, a phoneme model is typically comprised of three models which are tied together into one unit using an overarching Hidden Markov Model (HMM). HMMs provide an excellent framework to capture the variations in the phoneme realisations and durations.

The standard approach to train the HMMs is based on Expectation Maximisation using maximum likelihood criterion as proposed in [Dempster et al., 1977]. More recently, discriminative training criteria using Maximum Mutual Information [Normandin, 1991] and minimum bayes' risk based Minimum Phone Error [Povey and Woodland, 2002] have been proposed. In this thesis, we use the maximum likelihood criterion to train the models which will be discussed in more detail in the following sections.

3.1.2.1 Hidden Markov models for acoustic modeling

The basic theory of Hidden Markov models was proposed in the 1960s [Baum and Petrie, 1966]. It is essentially a Markov model in which the state sequence is not observable. It provides a good framework to model an observable time series in which the underlying system generating the observations can be assumed to lie in a finite set of states.

Some of the earliest acoustic models using HMMs were built at CMU [Baker, 1975] and IBM [Jelinek, 1976]. Due to the elegant framework they provide, HMMs have been adopted as a standard modeling technique for acoustic models.

HMMs can have a variety of configurations based on the allowed transitions between states. In speech processing, as explained above, we use one state to model a segment of a phoneme (either beginning, middle or ending). This enforces an ordering for state transitions in the HMM and hence a constrained left-to-right HMM is typically used.

The structure of a three state left to right HMM is shown in the Figure 3.7.

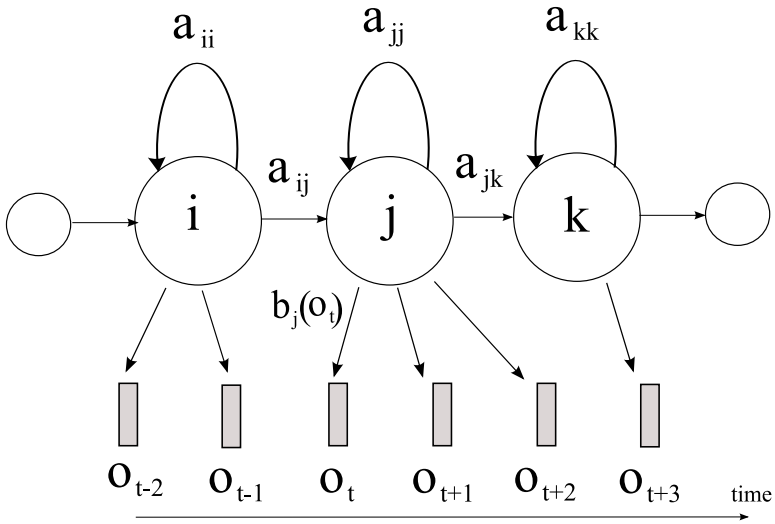


Figure 3.7: Three state left-right HMM

A HMM is specified by

1. The number of emitting states in the HMM 'N'. For ease of implementation, two dummy states one for 'Entry' and one for 'Exit' are appended which are used to concatenate HMMs together into larger context word-level or utterance level HMMs. Each emitting state contains a probability density function such as a multivariate Gaussian or a Gaussian mixture model.

2. Transition probabilities 'a_{ij}' which capture the probability of transition from state i to state j. i.e., $a_{ij} \quad P\,q_{\ t\ 1} \quad j|q_{\ t} \quad i$ where, $q_{\ t}$ indicates the state occupied at time t. a_{ij} terms must obey the condition $\sum_{j\ 1}^{N} a_{ij} \quad 1$. At every instance of time, there is a change in state from the current state to one of the states having a nonzero transition probability from the current state.

3. Emission probability densities $b_j\,o_t$ that capture the probability of observing feature vector o_t emitted from state j at time t. i.e., $b_j\,o_t \quad P\,o\,t\,|q_{\ t} \quad j$. The state output distribution modeled by a multivariate Gaussian distribution is given by

$$b_j\,o_t \qquad \frac{1}{\sqrt{2\pi\ ^{n}|\Sigma_j|}} e^{-\frac{1}{2}\ \varphi-\mu_j\ \Sigma_j^{-1}\ \varphi-\mu_j}$$

   where, μ_j and Σ_j are the mean and the covariance of the Gaussian respectively.

4. The initial state distributions 'π' that capture the probability of a state occurring at time $t \quad 1$, i.e., $\pi_{\ i} \quad P\,q_{\ 1} \quad i$

Thus a HMM is specified by the triplet $\lambda \quad A, B, \pi$.

In order to be able to use HMMs to represent the acoustics, the following assumptions are made

1. The speech waveform is stationary over segments of time i.e, over the duration of feature extraction window.
2. The probability of a state is only dependent on its previous state.
3. The transition between states is instantaneous.
4. The probability of an observation being generated is only dependent on the current state and independent of the previous states and observations.

$$P\ O|q_1, q_2, \ldots, q_T, \lambda \quad \prod_{t\ 1}^{T} P\ o_t|q_t, \lambda \tag{3.8}$$

Though not completely true, these assumptions allow the modeling of speech in the mathematical framework of HMMs. Elegant algorithms for training the HMMs and recognition of speech using those HMMs have been developed.

3.1.2.2 The three problems of Hidden Markov models

We briefly review the three central problems in the HMM theory. These are discussed in detail by Rabiner [1989].

Problem 1: Evaluation

Given a sequence of observations $O \quad o\ _1 o_2 .. o_T$, this problem deals with the computation of the likelihood of a given HMM (λ) generating O i.e., $P\ O|\lambda$. Since the state sequence $q \quad q\ _1 q_2 .. q_T$ is hidden, the likelihood is computed by marginalising over all possible state sequences.

$$P\ O|\lambda \quad \sum_{q} P\ O|q, \lambda\ P\ q|\lambda \tag{3.9}$$

where,

$$P\ O|q, \lambda \quad b\ _{q_1}\ o_1\ b_{q_2}\ o_2\ \ldots b_{q_T}\ o_T \tag{3.10}$$

$$P\ q|\lambda \quad \pi\ _{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \tag{3.11}$$

If there are N states in the HMM, the number of possible state sequences in T time steps is N^T, and hence the evaluation of the likelihood in equation 3.9 using the naive approach has complexity of $O\ TN^T$ which is computationally infeasible.

In order to make this calculation tractable, taking advantage of the Markovian assumption, recursive forward or backward procedures are used. Variables α and β are introduced in these two procedures respectively, which are used to accumulate the statistics as explained below and hence avoid the need to replicate calculations.

Forward procedure

The forward variable $\alpha_t\ i$ is defined as

$$\alpha_t\ i \quad P\ o\ \ _1, o_2, \ldots, o_t, q_t \quad i|\lambda \tag{3.12}$$

i.e., the probability of observing the partial sequence $o_1..o_t$ and occupying state i at time t.

Using induction, $\alpha_t\ i$ can be computed as follows:

1.
$$\alpha_1\ i \quad \pi\ \ _i b_i\ o_1\ , \quad 1 \leq i \leq N \tag{3.13}$$

2.
$$\alpha_t\ \ _1\ j \qquad \sum_{i\ \ 1}^{N} \alpha_t\ i\ a_{\ ij} \quad b_j\ o_t\ \ _1$$
$$\text{for} \quad t \quad 1, 2, \ldots, T-1, \quad 1 \leq j \leq N \tag{3.14}$$

3. The likelihood $P\ O|\lambda$ can be computed by marginalising the α variables at time T over all the states.
$$P\ O|\lambda \qquad \sum_{i\ \ 1}^{N} \alpha_T\ i \tag{3.15}$$

With this approach the computational complexity is of the order $O\ N^2 T$ which is a significant improvement over the naive approach.

Backward procedure

Similar to the forward procedure, $P\ O|\lambda$ can also computed using a backward procedure using a variable $\beta_t\ i$ which is defined as the the probability of observing the sequence from time instance $t \quad 1$ to T given the occupancy of state i of the model λ at time t.

$$\beta_t\ i \quad P\ o\ \ _{t\ \ 1}, o_{t\ \ 2}, \ldots, o_T|q_t \quad i, \lambda \tag{3.16}$$

The recursive formulation for likelihood computation using the backward procedure is as follows:

1.

$$\beta_T\ i \quad 1, \quad 1 \leq i \leq N \tag{3.17}$$

2.

$$\beta_t\ i \qquad \sum_{j\ 1}^{N} a_{ij} b_j\ o_{t\ \ 1}\ \beta_{t\ \ 1}\ j$$
$$\text{for} \quad t \quad T-1, T-2, \ldots, 1 \quad 1 \leq i \leq N \tag{3.18}$$

3. And

$$P\ O|\lambda \qquad \sum_{i\ 1}^{N} \pi_i b_i\ o_1\ \beta_1\ i \tag{3.19}$$

The computational complexity using the backward procedure is also of the order $O\ N^2 T$. Hence both the procedures are equally efficient to solve the evaluation problem.

Problem 2: Decoding

This is the problem of identifying the state sequence that maximises the likelihood of an observation sequence given the model i.e,

$$\arg\max_q P\ O, q|\lambda \qquad where \quad q \quad Q \tag{3.20}$$

This search problem can be efficiently solved using Viterbi decoding [Viterbi, 1967]. It is a dynamic programming approach which builds the complete solution from optimal sub solutions.

Let $\delta_t\ i$ be the likelihood of the partial observation sequence $o_{\ 1}..o_t$ generated by the best state sequence ending in state i at time t.

$$\delta_t\ i \quad \max_q \quad P\ q_1, q_2 \ldots q_t \quad i, o_{\ \ 1}, o_2 \ldots o_t|\lambda \tag{3.21}$$

By recurrent property, the value of δ at time $t \quad 1$ for state j can be computed by:

$$\delta_{t\ \ 1}\ j \quad \max_i \quad \delta_t\ i\ a_{\ ij}\ \ b_j\ o_t \tag{3.22}$$

As the Viterbi algorithm builds the solution step by step, it is also essential to keep track of the best predecessor state at each time instant. This information is stored in $\psi_t\ i$ which denotes the best preceding state for current state i at time t. The algorithm to find the optimal state sequence is as follows:

1. *Initialisation*

$$\delta_1(i) = \pi_i b_i(o_1) \tag{3.23}$$

$$\psi_1(i) = 0 \tag{3.24}$$

$$\text{for} \quad 1 \leq i \leq N$$

2. *Build optimal sub solutions iteratively*

   For time instances $t = 2$ to $T-1$, update the δs and ψs for each state j $(1 \leq j \leq N)$ using:

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \tag{3.25}$$

$$\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i) a_{ij}\right] \tag{3.26}$$

$$\text{for} \quad 1 \leq i \leq N$$

3. *Termination*

   Find the state at time T that has the maximum cumulative likelihood score.

$$p = \max_i \left[\delta_T(i)\right] \tag{3.27}$$

$$q_T = \arg\max_i \left[\delta_T(i)\right] \tag{3.28}$$

$$\text{for} \quad 1 \leq i \leq N$$

4. *Backtrack*

   Having found the best state q_T at time T, the best state sequence explaining the observation data O can be found by backtracking the states stored in ψ variable.

$$q_t = \psi_{t+1}(q_{t+1}) \quad \text{for} \quad t = T-1, T-2, \ldots, 1 \tag{3.29}$$

Problem 3: Learning

This is the problem of training the models from the observation sequences such that the models generalise well to unseen data of similar nature i.e., find the model parameters $\lambda(A, B, \pi)$ that maximise $P(O|\lambda)$.

The procedure employed in training the models is to start with an initial model λ_0 and update the parameters iteratively until the difference in the likelihood $P(O|\lambda)$ between two successive iterations becomes negligible.

The initial models are usually built using either

1. a segmental K Means approach, where the observation vectors are split equally among all the states in the HMM and the parameters estimated, or

2. a flat start approach where all the observation vectors are used to estimate the parameters of one global model which is then used as the seed model for all the units.

Starting from such initial models, the parameters are then updated iteratively in a maximum likelihood sense using Baum-Welch re-estimation [Baum et al., 1970] process.

Let $\gamma_t\ i$ be the probability of occupying state i at time t given the observation sequence O and the model λ.

$$\gamma_t\ i \quad p\ q \quad _t \quad i|O,\lambda \qquad \frac{P\ q_t \quad i,O|\lambda}{P\ O|\lambda} \qquad \frac{\alpha_t\ i\ \beta_t\ i}{\sum_{i \quad 1}^{N} \alpha_t\ i\ \beta_t\ i} \tag{3.30}$$

Let $\xi_t\ i,j$ be the joint probability of occupying state i at time t and occupying state j at time $t \quad 1$ given the observation sequence O and the model λ.

$$\xi_t\ i,j \qquad P\ q \quad _t \quad i,q \quad _{t \quad 1} \quad j|O,\lambda \qquad \frac{P\ q_t \quad i,q \quad _{t \quad 1} \quad j,O|\lambda}{P\ O|\lambda} \tag{3.31}$$

$$\frac{\alpha_t\ i\ a\ _{ij} b_j\ o_{t \quad 1}\ \beta_{t \quad 1}\ j}{P\ O|\lambda} \tag{3.32}$$

Summing $\gamma_t\ i$ from $t \quad 1$ to T gives the expected number of times state i is occupied and summing it over $t \quad 1$ to $T-1$ gives the expected number of times there is a transition from state i. Similarly summing $\xi_t\ i,j$ over $t \quad 1$ to $T-1$ gives the expected number of transitions from state i to state j.

The Baum-Welch re-estimation formulae to update the model parameters assuming Gaussian densities are:

$$\widehat{\pi}_i \qquad \gamma\ _1\ i \tag{3.33}$$

$$\widehat{a}_{ij} \qquad \frac{\sum_{t \quad 1}^{T-1} \xi_t\ i,j}{\sum_{t \quad 1}^{T-1} \gamma_t\ i} \tag{3.34}$$

$$\widehat{\mu}_i \qquad \frac{\sum_{t \quad 1}^{T} \gamma_t\ i\ o_t}{\sum_{t \quad 1}^{T} \gamma_t\ i} \tag{3.35}$$

$$\widehat{\Sigma}_i \qquad \frac{\sum_{t \quad 1}^{T} \gamma_t\ i\ \ o_t - \mu_j\ \ o_t - \mu_j}{\sum_{t \quad 1}^{T} \gamma_t\ i} \tag{3.36}$$

3.1.2.3 Extensions

Gaussian mixture models

Using single Gaussians to model each state of the HMM would imply an assumption that all the data used to model a state is unimodal. However in a large dimensional space, this assumption is not appropriate. So modifying the assumption of the underlying data distribution from unimodal to multimodal, single Gaussians are replaced by Gaussian Mixture Models (GMM) and the emission probability density of state j is given by:

$$b_j\ o_t \qquad \sum_{m\ \ 1}^{M_j} w_{jm}\mathcal{N}\ o_t;\mu_{jm},\Sigma_{jm} \tag{3.37}$$

where M_j denotes the number of mixture components in the GMM modeling state j and w_{jm}, μ_{jm} and Σ_{jm} denote the weight, mean and variance of the m^{th} component in state j.

The re-estimation formulae for the HMM parameters are modified as follows:

The term $\gamma_t\ i$ is modified to $\gamma_{\ t}\ i,k$ which accounts for the probability of occupying mixture component k in state i at time t given the observation sequence O and the model λ.

$$\gamma_t\ i,k \qquad \frac{\alpha_t\ i\ \beta_{\ t}\ i}{\sum_{i\ \ 1}^{N}\alpha_t\ i\ \beta_{\ t}\ i} \qquad \frac{w_{ik}\mathcal{N}\ o_t;\mu_{ik},\Sigma_{ik}}{\sum_{m\ \ 1}^{M} w_{im}\mathcal{N}\ o_t;\mu_{im},\Sigma_{im}} \tag{3.38}$$

Using $\gamma_t\ i,k$ the formulae for computation of the weight, mean and variance of each mixture component are given by:

$$\widehat{w}_{ik} \qquad \frac{\sum_{t\ \ 1}^{T}\gamma_t\ i,k}{\sum_{t\ \ 1}^{T}\sum_{m\ \ 1}^{M}\gamma_t\ i,m} \tag{3.39}$$

$$\widehat{\mu}_{ik} \qquad \frac{\sum_{t\ \ 1}^{T}\gamma_t\ i,k\ o_{\ t}}{\sum_{t\ \ 1}^{T}\gamma_t\ i,k} \tag{3.40}$$

$$\widehat{\Sigma}_{ik} \qquad \frac{\sum_{t\ \ 1}^{T}\gamma_t\ i,k\ \ o_{\ t}-\widehat{\mu}_{ik}\ \ o_{\ t}-\widehat{\mu}_{ik}}{\sum_{t\ \ 1}^{T}\gamma_t\ i,k} \tag{3.41}$$

While training the GMM-HMM based systems, the number of components in the GMM cannot be determined in advance. The usual approach is to start the training process with one Gaussian per state, and then repeatedly train and split the available mixture components till the likelihood of the model is maximised for a small development set disjoint from the training set. Mixture sizes of 16 to 32 components are typically used in most of the state-of-the-art acoustic models [Hain et al., 2008].

Triphone context

To overcome the variability in the phoneme pronunciations introduced due to co-articulation effect with adjacent phonemes, triphone models are usually preferred in large vocabulary systems. Hence for each phoneme, instead of having a single 3 state HMM, multiple HMMs with different left and right contexts are trained. Triphones are usually represented as $L-P\ \ R$ where L and R are the left and right context respectively for the phoneme P.

Example: HMM for the word '*monotonic*' would be comprised of concatenation of following monophone HMMs: `m aa n ax t aa n ih k`
Using Triphone HMMs it would be comprised of:
`sil-m+aa m-aa+n aa-n+ax ax-t+aa aa-n+ih ih-k+sil`
Here two different HMMs are used for 'aa' depending on the phonetic context to the left and right.

Triphones can have word-internal or crossword context [Young et al., 2006]. In word internal triphones, the left and right context are limited to the word boundaries in the transcript. In such cases, the boundary phonemes have only one of the left or right context as shown in the example below. Crossword context triphones on the other hand use context from adjacent words for boundary phonemes. In all the experiments presented in this thesis, unless otherwise specified, crossword context dependent triphones have been used.

Transcript: This is speech
Monophone sequence: `sil th ih s sp ih s sp s p iy ch sil`
Word Internal context dependent triphones: `sil th+ih th-ih+s ih-s sp ih+s ih-s sp s+p s-p+iy p-iy+ch iy-ch sil`
Crossword context dependent triphones: `sil sil-th+ih th-ih+s ih-s+ih sp s-ih+s ih-s+s sp s-s+p s-p+iy p-iy+ch iy-ch+sil sil`

State tying

When we use triphones instead of monophones, the total number of HMMs increase cubically. For instance the number of monophones in the CMU phoneme set that we use in the experiments is 41 and the number of triphones covering all the left-right contexts is 68921. This increase in the number of models leads to data sparsity

problem during the training phase. In fact several of the triphones may not appear even once in the training data.

However, several states of all the triphones are acoustically close to each other and hence can be tied together. For eg., it is desirable that the 2^{nd} and 3^{rd} states of the two triphones 'k-aa+r' and 'b-aa+r' be tied. This allows the training data from all the tied states to be pooled together to form a larger training set. It is not desirable though, to tie the states using hand written rules. Several researchers have worked on this problem in the early nineties.

One way to automate the state tying is by data driven clustering of the states using either a top-down or bottom-up approach. Hwang and Huang [1992] propose an agglomerative clustering approach where by starting with one cluster per state, desired number of clusters are generated by repeated merging. One disadvantage of this approach is that it cannot account for unseen triphones in the training data.

The other approach using phonetic decision trees [Bahl et al., 1991; Young et al., 1994] provides a mechanism to cluster even unseen triphones. A phonetic decision tree is a binary tree with a yes/no phonetic question associated with each node of the tree. The phonetic questions for example take the form 'Is the right context a Fricative?', 'Is the phoneme an affricate?' and so on. Each phonetic state trickles down from the root node to one of the leaf nodes depending on the answer at each intermediate node's phonetic question. All the states arriving at the same leaf node are tied together. Even unseen phonemes can be clustered in this manner. The decision tree itself is built from a predefined set of questions in a top down manner. The question associated with each node is chosen such that the likelihood of the resulting tied states is maximised for the training data. Starting from the root node, the tree grows associating a question to each node until the gain in likelihood from further splits falls below a predefined threshold.

3.1.3 Language models

Language models contain information about the allowable word sequences. They help in limiting the word-search space for the recogniser and improve the accuracy and reduce the computational load. Language models may take the form of word networks which limit the set of words that can follow a word by the rules of the graph or statistical models such as N-Gram [Ney et al., 1994], that associate probabilities to word sequences.

Word network lattices are preferred in small scale ASR systems. In systems where

the user's input utterance is limited to a few words and where the context is predefined, use of word network lattices can lead to very high recognition accuracies. Grammars are usually hand written in Extended Backus-Naur Form (EBNF). Since the grammar sizes are very small, they can be loaded and freed from the system memory at run time with negligible delays. This facilitates in changing the grammar depending on the context or the state of the spoken dialogue system. For this reason they are widely used in the commercial spoken dialogue systems. Figure 3.8 shows an example of a grammar lattice that can be used in an appointment scheduling dialogue system when the user needs to select a suitable appointment session during the week.

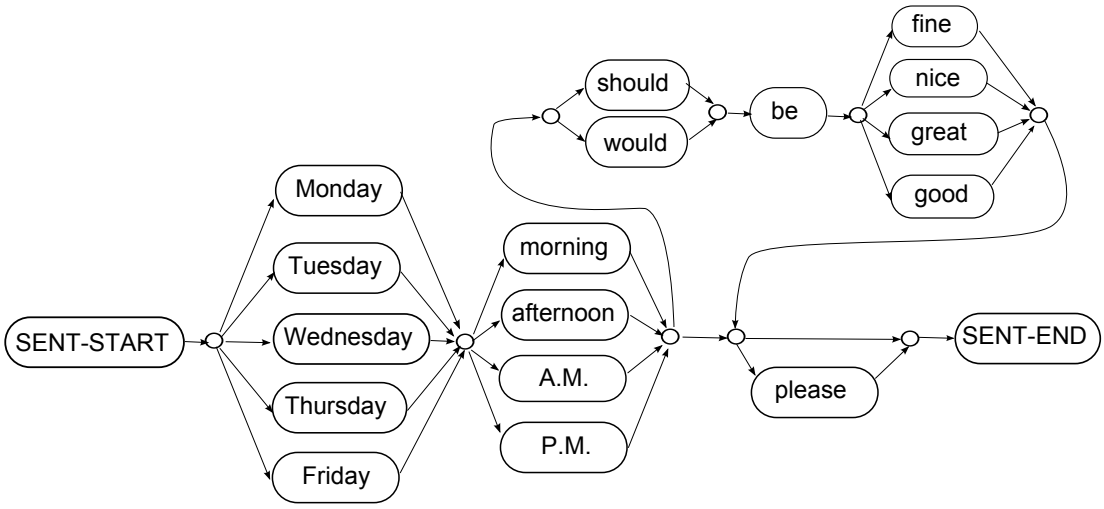


Figure 3.8: Example of a word network lattice

In large vocabulary systems that allow continuous speech from the user, hand written word lattices are infeasible. In such systems, statistical n-gram language models covering a wide range of words are used. Such language models are usually trained on large text corpora such as newspaper articles, text generated from web crawling and text collected specifically to represent well the domain of usage.

A statistical language model is used to compute the probability of a sequence of words occurring together in the language. Decomposing the joint probability as the product of conditional probabilities, the probability of the word sequence is given by:

$$P(w_1, w_2, \ldots, w_m) \qquad P(w_1)\,P(w_2|w_1)\,P(w_3|w_1, w_2)\ldots P(w_m|w_1 \ldots w_{m-1})$$
$$\prod_{i=1}^{m} P(w_i|w_1, \ldots, w_{i-1}) \tag{3.42}$$

It is infeasible to store the statistics for every word given all possible context lengths. Hence in practice n-gram language models are used, where the statistics for a word

given a word sequence of $n-1$ predecessors are stored. This constraint on the history amounts to the following assumption

$$P\ w_m|w_1, w_2, \ldots w_{m-1} \quad P\ w_{\ m}|w_{m-n\ \ 1}, \ldots w_{m-1} \tag{3.43}$$

In ideal conditions, a large value of n would be preferred, to have the best approximation, but due to the storage limitations and the sparsity of training data, n values of 1 to 3 are typically used. With the advent of cloud computing infrastructure and trillions of documents of data available on the world wide web, researchers have only recently begun experimenting with higher order language models such as 5-gram models [Brants and Franz, 2006].

Using word n-grams, the probability of a word sequence modifies to

$$P\ w_1, w_2, \ldots, w_m \quad \prod_{i\ \ 1}^{m} P\ w_i|w_{i-n\ \ 1}, \ldots, w_{i-1} \tag{3.44}$$

The n-gram sequence probabilities are estimated in a maximum likelihood sense by counting the word sequence instances in the training text corpora.

$$P\ w_m|w_{m-n\ \ 1} \ldots, w_{m-1} \quad \frac{c\ w_{m-n\ \ 1} \ldots w_m}{c\ w_{m-n\ \ 1} \ldots, w_{m-1}} \tag{3.45}$$

where, c represents the count of word sequence in the training corpora.

This naive approach to compute the maximum likelihood estimates of the n-gram probabilities runs into problems with data sparsity. The n-grams not observed in the training set are assigned zero probability which is not desirable as such word sequences may appear in the test set. This is a common recurring problem in statistical modeling usually solved by some kind of a smoothing technique. Smoothing for language models [Chen and Goodman, 1996] are based on three main ideas

1. Discounting, where some probability mass from high frequency word types is reassigned to those with near zero frequency.

2. Backoff, where for an unseen n-gram, the conditional probability of word given its history is approximated by backing off to the conditional probability of that word given a shorter context.

3. Interpolation, where the conditional probability for a higher order n-gram is computed as a linear combination of the probability estimates of the lower order (shorter context) n-grams.

Smoothing

One of the earliest proposed smoothing method ***Good-Turing discounting*** [Good, 1953] assigns some of the probability mass of n-grams occurring c 1 times to n-grams occurring c times in the training corpus. The adjusted counts for all the n-grams with a count of c is given by

$$c \quad c \ 1 \quad \frac{N_{c \ 1}}{N_c} \tag{3.46}$$

where, N_c denotes the number of n-grams with count c.

For a sufficiently large corpus, $N_{c \ 1}$ is usually less than N_c and hence the adjusted counts are less than the actual counts leaving some probability mass to be assigned to unseen n-grams. The adjusted counts in equation 3.46 are however unreliable at higher values of c where N_c values would be zero or near zero.

A modified equation for the discounted counts was proposed in [Katz, 1987] which is used in conjunction with a backoff technique.

$$\begin{aligned} c \quad & c \quad \text{for,} \quad c \quad k \\ & \frac{c \ 1 \quad \frac{N_{c \ 1}}{N_c} - c\frac{k \ 1 \ kN_1}{N_1}}{1 - \frac{k \ 1 \ kN_1}{N_1}} \quad \text{for,} \quad 1 \leq c \leq k \end{aligned} \tag{3.47}$$

k is a threshold that can be set to zero or determined empirically.

In ***Katz backoff smoothing***, when the n-gram counts are zero, the model backs off to a lower order model with a non-zero count. In order to maintain correct probability distributions, n-grams with non-zero counts are discounted such that when a probability for a zero count n-gram is assigned from a lower order, the sum of the probability for a given word w_n given all contexts, sums to one.

$$P_{bo}\ w_m | w_{m-n \ 1} \ldots w_{m-1} \qquad d \ _{m-n \ 1 \ldots m-1} \frac{C\ w_{m-n \ 1} \ldots w_m}{c\ w_{m-n \ 1} \ldots w_{m-1}}$$

$$\text{if} \quad C\ w_{m-n \ 1} \ldots w_m \quad k \tag{3.48}$$

$$\alpha \ _{w_{m-n \ 1} \ldots w_{m-1}} P_{bo}\ w_m | w_{m-n \ 2} \ldots w_{m-1}$$

$$\text{otherwise} \tag{3.49}$$

Equation 3.48 discounts the n-grams occurring with a count greater than certain threshold. The discounting factor d can be found from modified Good-Turing estimates in equation 3.47

$$d \quad \frac{c}{c} \tag{3.50}$$

To compute the α values, the total left over probability mass after discounting the n-grams is first accumulated as β and this is redistributed equally among all the n-grams with zero count.

$$\beta_{w_{m-n\ 1}\ldots w_{m-1}} \quad 1 - \sum_{w_m : C\ w_{m-n\ 1}\ldots w_m\ \ 0} d_{w_{m-n\ 1}\ldots w_{m-1}} \frac{C\ w_{m-n\ 1}\ldots w_m}{C\ w_{m-n\ 1}\ldots w_{m-1}} \tag{3.51}$$

$$\alpha_{w_{m-n\ 1}\ldots w_{m-1}} \quad \frac{\beta_{w_{m-n\ 1}\ldots w_{m-1}}}{\sum_{w_m : C\ w_{m-n\ 1}\ldots w_m\ \ 0} P_{bo}\ w_m | w_{m-n\ 2}\ldots w_{m-1}} \tag{3.52}$$

Among the smoothing methods commonly used, ***Kneser-Ney*** Smoothing [Ney et al., 1994; Kneser and Ney, 1995], was shown to give the best performance in terms of ASR WERs [Chen and Goodman, 1996]. Kneser-Ney smoothing is based on a simpler approach called ***Absolute discounting***, where the non-zero n-grams counts are discounted by an absolute value D. The discounted probabilities with absolute discounting for a bi-gram are:

$$P_{absolute}\ w_m | w_{w_{m-1}} \quad \begin{cases} \dfrac{C\ w_{m-1}, w_m\ - D}{C\ w_{m-1}} & \text{if} \quad C\ w_{m-1}, w_m \quad 0 \\ \alpha\ w_m\ P_{ML}\ w_m & \text{otherwise} \end{cases} \tag{3.53}$$

Kneser-Ney smoothing uses a slightly different approach to compute the unigram probabilities. Instead of counting the number of times a word occurs in the corpus, the number of different contexts in which a word appears is counted.

$$C_{KN}\ w_m \quad |w\ \ _{m-1} : C\ w_{m-1}, w_m \quad 0\ | \tag{3.54}$$

The idea behind the use of such a count is that those words that have appeared in more contexts in the training set are more likely to appear in unseen contexts as well. Using these counts, the estimated probability is termed as 'Continuation Probability' P_{cont} [Jurafsky and Martin, 2008].

$$P_{cont}\ w_m \quad \frac{|w_{m-1} : C\ w_{m-1}, w_m \quad 0\ |}{\sum_{w_i} |w_{i-1} : C\ w_{i-1}, w_i \quad 0\ |} \tag{3.55}$$

For Knerser-Ney backoff smoothing, equation 3.53 is modified using the continuation probabilities as follows:

$$P_{KN}\ w_m | w_{w_{m-1}} \quad \begin{cases} \dfrac{C\ w_{m-1}, w_m\ - D}{C\ w_{m-1}} & \text{if} \quad C\ w_{m-1}, w_m \quad 0 \\ \alpha\ w_m\ \dfrac{|w_{m-1} : C\ w_{m-1}, w_m \quad 0\ |}{\sum_{w_i} |w_{i-1} : C\ w_{i-1}, w_i \quad 0\ |} & \text{otherwise} \end{cases} \tag{3.56}$$

The ***Deleted interpolation*** smoothing [Jelinek and Mercer, 1980] interpolates the maximum likelihood estimates of a word with estimates from shorter contexts. For instance, for a trigram language model, the adjusted probability estimate is given by:

$$P_{DI}\ w_n|w_{n-1},w_{n-2} \quad \lambda\ _1 P_{ML}\ w_n|w_{n-1},w_{n-2}\ \ \lambda\ _2 P_{ML}\ w_n|w_{n-1} \quad \lambda\ _3 P_{ML}\ w_n \tag{3.57}$$

where, P_{ML} is the maximum likelihood estimate as given in equation 3.45 and, $\sum_i \lambda_i \quad 1$

Equation 3.58 can be expressed in an elegant recursive formulation after Brown et al. [1992]:

$$P_{DI}\ w_n|w_{m-n\ \ 1}\ldots w_{m-1} \quad \lambda\ _{m-n\ \ 1} P_{ML}\ w_n|w_{m-n\ \ 1}\ldots w_{m-1} \quad 1-\lambda\ _{m-n\ \ 1}\ P_{DI}\ w_n|w_{m-n\ \ 2}\ldots w_{m-1} \tag{3.58}$$

Here the n^{th} order smoothed estimate is a linear combination of n^{th} order maximum likelihood estimate and the $n-1\ ^{th}$ order smoothed estimate. The recursion can be terminated by approximating the smoothed first order estimate to be equal to the maximum likelihood estimate. The λ values are computed using EM algorithm[Dempster et al., 1977] such that the probability of some held out development set is maximised.

Perplexity

Given a statistical language model, it is of interest to evaluate how well it models the language in the domain of interest. A performance measure metric is also desirable to compare the performance of different language models for a given test set.

Based on principles from Shannon's Information Theory [Shannon, 1948], entropy (H) of a language generating source measures the amount of non-redundant information contained in a word sequence in that language.

$$H \quad -\lim_{m\ \infty}\frac{1}{m}\sum_{w_1\ldots w_m} P\ w_{1}\ldots w_m\ \log_2 P\ w_1\ldots w_m \tag{3.59}$$

where, P is the true probability of the word sequence.

Assuming ergodicity and sufficiently large value of m, entropy can be approximated as

$$\widehat{H} \quad -\frac{1}{m}\log_2 P\ w_1\ldots w_m \tag{3.60}$$

Instead of directly using entropy, a related measure called perplexity (PP) is generally used.

$$PP \quad 2^{\ \widehat{H}} \tag{3.61}$$

If the probability of the word sequence generated by the language model is given by $\tilde{P}$, then the perplexity is given by

$$PP \quad \tilde{P}\, w_1 \ldots w_m \; ^{-\frac{1}{m}} \tag{3.62}$$

When the language model is a good fit, it assigns high probabilities to unseen test sequences and thereby lower perplexities. Though perplexity is a good measure to compare language models, it is found to be weakly correlated to the measure of interest in ASR viz., Word Error Rate.

3.1.4 Lexicon

A lexicon (dictionary) is a collection of all the words (vocabulary) used in the ASR system and provides a map between the units that are represented in the acoustic models and the words present in the language model. In a small vocabulary task using word based acoustic models, the lexicon would be a simple one-to-one mapping between words and the symbolic representation of the acoustic models. In a large vocabulary task, the acoustic models are trained as sub-word units such as phonemes and the lexicon provides a map between these sub-word units and the words.

Example:
ABOARD ax b ao r d
ABOLISH ax b aa l ih sh
ABSENT ax b s ax n t
ABSORB ax b z ao r b
..

Alternate pronunciations of words can be encoded in the lexicon and thus provide more flexibility to the ASR system to deal with dialects. The lexicon is typically built apriori using rules of pronunciation for that language. For certain outliers and difficult words, the pronunciation is hand coded.

The vocabulary, in effect sets the possible words that can be decoded by the ASR. Any word spoken by the user that is outside the vocabulary is mapped to a close word in the vocabulary and is one of the main sources of error and thereby poor recognition accuracies in ASR systems. Hence it is imperative to carefully cover the all possible word tokens for the domain of usage, in the vocabulary.

3.1.5 Decoder

The role of the decoders is to combine the acoustic model and language model scores, searching through all possible word sequences and output the best hypothesis for the spoken utterance transcription. Most state-of-the-art decoders use a tree structure or a finite state transducer (FST) to represent the search space. An FST used in ASR comprises of nodes corresponding to acoustic models and nodes that correspond to word endings. As the search progresses through a node, depending on the type of node, either the acoustic model score or the language model score is added to that search path.

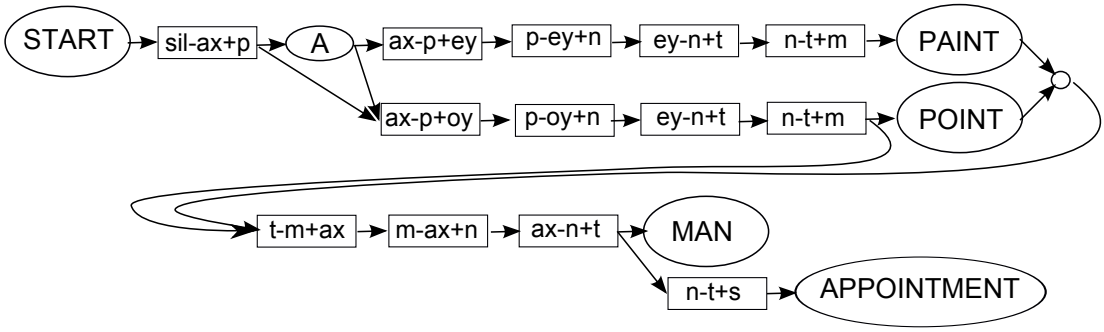


Figure 3.9: Example of a small segment of a finite state network

Usually an FST is built statically and preloaded into the memory before the hypothesis search begins. However in large vocabulary systems, the size of the FST can be extremely large and the use of dynamic network expansion approaches have been suggested. Efficient wighted finite state transducer algorithms have been proposed [Mohri et al., 1996] to construct compact decoding networks from the transducers of individual components in the ASR system.

The search itself can progress time synchronously in a breadth first search manner keeping track of the best partial state sequences at each time instant. The Viterbi algorithm discussed in section 3.1.2.2 is an example of this type of search. Efficient best-first search algorithms such as Stack decoders and A decoders have also been quite well researched. These algorithms are time asynchronous. The key idea in these algorithms is to maintain a priority queue of partial sequences where each sequence maintains a score based on the acoustic and language model probabilities. The sequence iteratively chooses the sequence with the best score and determines the best word to post fix to the chosen sequence. The extended sequence is reentered into the queue. The computational complexity is drastically reduced since the algorithm fo-

cuses only the most probable candidates. Such efficiency in computational complexity can also be achieved in Viterbi style decoding by path pruning.

The decoding can also be done with multiple passes over the search network [Austin et al., 1991]. In such methods, usually a simple language model is used in the first pass, generating a list of N-best hypothesis. This set of hypothesis is then re-scored using higher order language models and possibly more sophisticated acoustic models. Multi-pass algorithms are have been shown to give better WERs at the expense of the the extra computational time.

In most of the decoding experiments in this thesis, the HTK tools HVite and HDecode have been used, which are based on Viterbi decoding. By treating the path extension in the search problem as a token passing instance, the search problem is posed as a token passing algorithm [Young et al., 1989]. This algorithm starts with a single token in the start node. As acoustic features are input to the system, tokens are passed to connected nodes in the network. When a token is passed from one node to another, the score associated with the token is updated with the probability associated with the new node. At each node, only the token with the highest score is retained and all other tokens are discarded. A word link record (WLR) which maintains the word sequence a token has traversed, is also maintained with each active token. In order to make the search computationally tractable, tokens with a score lower than the token with the highest score, by a certain threshold called beam width are pruned and their WLRs deleted. As a result only a small fraction of the tokens are active at any time instance and the network is expanded dynamically only for these paths.

The decoding setup in this thesis with the JNAS corpus uses the Julius decoder [Lee et al., 2001] in a two pass decoding mode. In the first pass, a forward frame synchronous beam search is used to generate a word trellis structure. In the second pass, a reverse search is performed on this word trellis using a stack decoding algorithm.

3.1.6 Performance measures

The performance of an ASR system is typically measured in terms of errors made by the system. Using dynamic programming based string alignment, the hypothesis generated from the decoder is aligned to the true transcription. The number of Hits (H), Substitutions (S), Deletions (D) and Insertions (I) are then computed.

The percentage of correct recognitions is given by

$$Correct\ \% \quad \frac{H}{H\ \ S\ \ D} \quad 100\% \tag{3.63}$$

The measure Word Error Rate (WER) is more commonly used

$$WER\ \% \quad \frac{S\ \ D\ \ I}{H\ \ S\ \ D} \quad 100\% \tag{3.64}$$

3.2 Normalisation approaches in acoustic space

An inherent problem in acoustic modeling, as in any machine learning problem, is the mismatch between the training set and the test set. This mismatch could be due to inadequate representation of test set speaker characteristics in the training set or due to the differences in environmental conditions such as recording setup, channel conditions and the ambient background noise. There are two main approaches to overcome this problem

1. Normalisation, where the acoustic features or the models themselves are normalised to remove undesirable variations.

2. Adaptation, where the features or models are adapted such that the differences due to mismatch conditions are well captured.

In this section, some of the widely used normalisation techniques are discussed.

3.2.1 Cepstral mean and variance normalisation

One of the common problems in speech corpora is the difference in channel (microphone) characteristics between various sessions. Cepstral Mean Normalisation (CMN) can effectively reduce the variations due to channel distortions. As explained in section 3.1.1.5, the channel characteristics which get convolved with the signal in time domain, appear as an addition in the cepstral domain. When the cepstral features are averaged over time, the mean represents the channel characteristics assuming that the channel is stationary. This mean is subtracted from the cepstral features to nullify the channel characteristics.

Cepstral Variance Normalisation (CVN) is typically used in conjunction with CMN. After mean normalisation, CVN involves scaling the feature vectors such that each feature in the vector has unit variance.

$$\mu_c \qquad \frac{1}{T}\sum_{t}^{T} c_t \tag{3.65}$$

$$\sigma_c^2 \qquad \frac{1}{T}\sum_{t}^{T} c_t^2 - \mu_t^2 \tag{3.66}$$

$$\widehat{c_t} \qquad \frac{c_t - \mu_c}{\sigma_c} \tag{3.67}$$

where c_t and $\widehat{c_t}$ are the original and CMN-CVN normalised cepstral features at time t.

CMN and CVN have been empirically shown to provide robustness to channel variations and white noise. In practical applications, instead of applying CMN-CVN over each utterance, they are often applied over longer segments of speech in which either the speaker or channel conditions are constant. In real time systems, the cepstral means and variances are computed as run time averages.

3.2.2 Vocal tract length normalisation

Vocal tract length shows a great degree of variability from speaker to speaker and more prominently between gender and age groups. As a result of this the formant peaks are different across speakers for the same spoken phoneme. The technique known as vocal tract length normalisation (VTLN) attempts to warp the frequency axis to compensate this difference.

Using a linear warping, significant improvements in WERs were achieved by Cohen et al. [1995]. The extent of the warping is determined by a warping factor α which is estimated for each speaker. To keep the warped frequencies bounded to the original frequency range, piecewise linear warping is usually applied with the boundary frequencies unwarped as shown in Figure 3.10

In various experiments in this thesis, warping as shown in Figure 3.10 (b) is used. The warping factor for each speaker is estimated in a maximum likelihood sense [Lee and Rose, 1996; Hain et al., 1999]. Through a linear search through various warping factors, the factor that maximises the likelihood of the acoustic model on some training data from the speaker is selected. Using a Brent search based on quadratic interpolation, the complexity of the search is substantially reduced. VTLN is applied in an iterative manner as described in Garau et al. [2005]

Training

1. Starting with non-normalised models, compute the warping factors for all

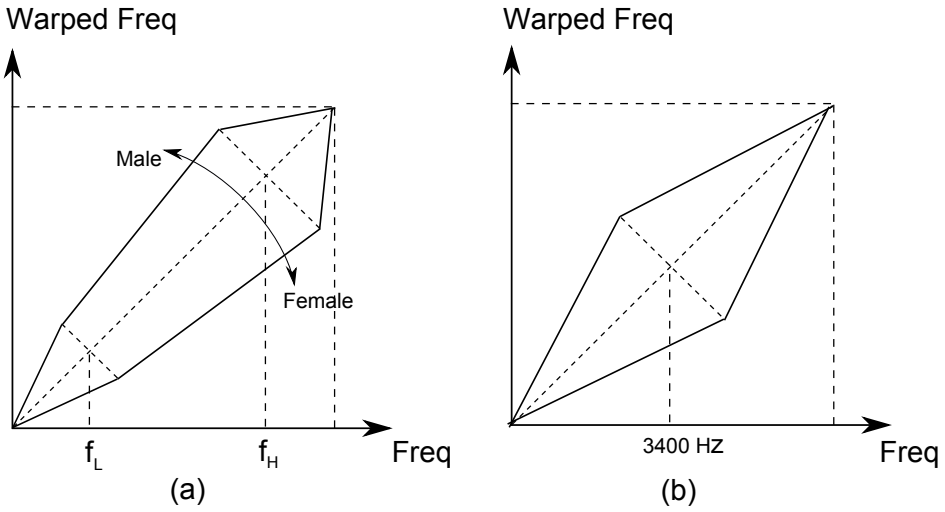


Figure 3.10: Piecewise linear warping in VTLN

the training set speakers. Using the warping factors, recompute the feature vectors for the entire training set.

2. Reestimate the parameters of the models with the new features with a few iterations of Baum-Welch algorithm.
3. Repeat steps 1 and 2 using the last retrained models until the increase in likelihood for some predefined development set stabilises.

Testing

1. For each test speaker, evaluate the warping factor using normalised models either with a small available development set, or by using the transcripts of the test set from one pass of decoding using the non-normalised model.
2. Normalise the acoustic features and decode with the normalised models.

Since the estimation is ML based, the warping factors are optimal in this framework where all other parameters are estimated in ML sense. The disadvantage of this approach is the high computational complexity due to the need to recode the data for various values of α.

The use of linear transforms to directly warp in the feature space has been explored by many researchers. McDonough et al. [1998] propose the use of a bilinear all pass transform for speaker normalisation. Pitz et al. [2001] explore the idea of applying linear transform after Cepstral extraction while Uebel and Woodland [1999] investigate the linear relationship between unwarped and warped MFCCs.

For VTLN to work reliably, speaker segmentation is essential. Such segmentation is naturally available in certain corpora such as telephone conversations while in scenarios such as meeting room conversations, a front end speaker diarization system is usually required to provide the same.

3.3 Adaptation approaches in acoustic space

Speaker independent acoustic models trained on hundreds of hours of speech from many speakers, generalise well for unseen speakers. However compared with speaker dependent models targeted towards a specific speaker, the WERs are high. But it is infeasible to train speaker dependent models for each speaker. Hence several approaches have been proposed to adapt the speaker independent models to a target speaker. All these approaches fall under the gamut of speaker adaptation. Speaker adaptation techniques have also been widely used for environment adaptation where there is a mismatch between train and test conditions.

Acoustic model adaptation attempts to modify the HMM parameters such that they resemble the target speech more closely. Several approaches to acoustic model adaptation have been proposed and these can be broadly classified into

Maximum likelihood based approaches.

Maximum-a posteriori (MAP) based approaches

Approaches based on speaker clustering.

3.3.1 Maximum likelihood adaptation

The underlying idea in ML based adaptation is to update the parameters of the acoustic models such that the likelihood of a set of target speech is maximised. Most of the popular approaches in this domain estimate a linear transform from the adaptation data to modify the HMM parameters.

On the face of it, estimating transformation matrices for each parameter of all the HMMs in the acoustic model seems infeasible due to the requirement of large amounts of training data. However, it has been shown that significant improvements in accuracies can be achieved by tying several states of the HMMs together and estimating one transform for all the tied parameters. This greatly reduces the number of parameters to be estimated and thus the need for large set of adaptation data. In fact, improvements

in WER can be achieved by estimating a single transform for all the model parameters thus making rapid adaptation possible even with a small amount of data.

3.3.1.1 MLLR

The core approach in this category is Maximum Likelihood Linear Regression (MLLR). In the basic MLLR [Leggetter and Woodland, 1995a] the means of the Gaussian components μ are updated using the following transformation

$$\bar{\mu} \quad A\mu \quad b \tag{3.68}$$

where A is an *n*x*n* regression matrix and b is an n-dimensional bias vector. A and b are computed using Expectation - Maximisation (E-M) algorithm [Dempster et al., 1977] such that the likelihood of the transformed models is maximised with respect to the adaptation data. This equation is more widely written in the form

$$\bar{\mu} \quad W \quad \widehat{\mu} \tag{3.69}$$

where W is an n x n 1 matrix and $\widehat{\mu}$ is the extended mean vector

$$\widehat{\mu}^T \quad 1 \quad \mu_1 \ldots \mu_n \tag{3.70}$$

Although a single global transform can be used for all the Gaussians, with availability of larger quantities of adaptation data more precise transforms can be computed that apply to a smaller number of Gaussians. One solution to achieve this is the regression class tree [Leggetter and Woodland, 1995b] where Gaussians that are close in acoustic space are clustered together and undergo the same transformation. Figure 3.11 (A) shows an example of a regression class tree.

The steps involved in applying MLLR based speaker adaptation are as follows:

1. Create the regression class tree based on acoustic distance between phonemes. The trees are usually built in a top-down approach using centroid splitting approach. Typically the Gaussian components in non-speech units such as *sil* and *sp* are tied together and form a child node of the root as shown in Figure 3.11 (B).

2. Accumulate the statistics for all the phonemes from the adaptation data with reference to the speaker independent models.

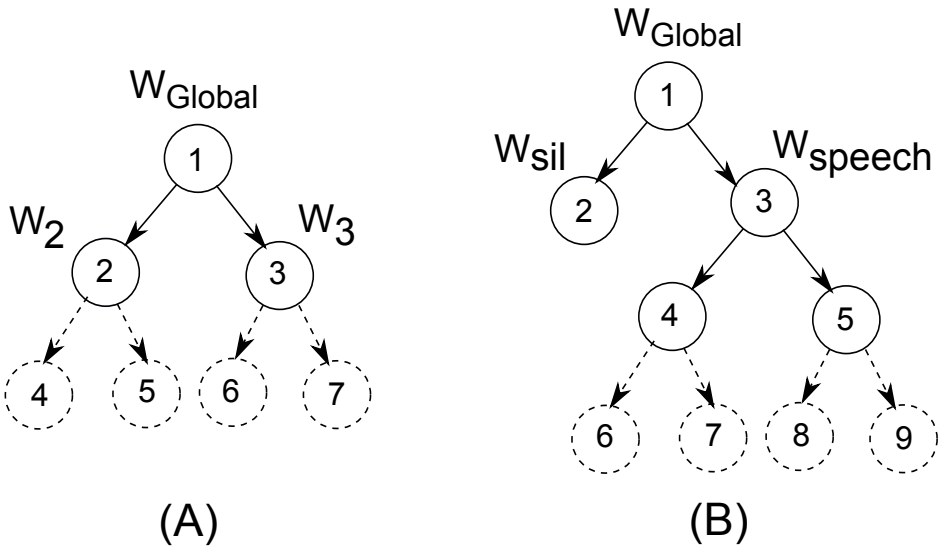


Figure 3.11: Regression trees

3. Using the statistics for all the phoneme states tied together at each node in the regression tree, estimate the transformation matrix for that node. If the accumulated statistics for a given node are not sufficient due to sparse data set, then the node borrows the transform from its parent node.

4. Decode the test utterances using models with parameters updated by the MLLR transforms.

To accumulate the statistics for each phoneme, the transcription for the adaptation utterances is essential. In the absence of transcripts, the accumulation is done in an unsupervised manner, where the utterances are first decoded using the speaker independent models and the decoder hypothesis is used as an approximate transcription.

Under the MLLR framework, Gaussian covariance matrix can also be adapted [Gales and Woodland, 1996; Gales, 1998]. The transforms to update the variances as proposed by Gales and Woodland [1996] take the form

$$\widehat{\Sigma} \quad LHL^{T} \tag{3.71}$$

where H is the linear transformation to be estimated and L is the Choleski factor of the covariance matrix Σ. Typically the transformation matrices for the means are first estimated and using the transformed means, the variance transformation matrices are estimated. Thus two different different matrices are required to transform the mean and variance of a single Gaussian and this approach is known as unconstrained adaptation.

Using MLLR for adaptation, significant improvements in large vocabulary continuous speech recognition have been obtained in several experiments. MLLR Transforms have more recently found application in speaker recognition applications too. In [Stolcke et al., 2005] a set of MLLR transforms for each speaker are used as representative vectors and Support Vector Machine classifiers are used to recognise the test speaker.

3.3.1.2 CMLLR (Constrained MLLR)

Unlike the standard MLLR, where independent transforms are computed for the Mean and Variance adaptation, the use of a single constrained transform was proposed by Digalakis et al. [1995] to update all the parameters of a single Gaussian. This variant of MLLR is known as Constrained MLLR (CMLLR).

The update equations in CMLLR are as follows:

$$\widehat{\mu} \quad A_c \mu - b_c \tag{3.72}$$

$$\widehat{\Sigma} \quad A_c^T \Sigma A_c \tag{3.73}$$

where, A_c and b_c are the constrained transform and the bias vector which need to be estimated in a maximum likelihood sense from the adaptation data.

Due to the constrained nature of the transforms, instead of transforming the models themselves, they can also be used to transform the observation vectors. Applying CMLLR, an observation vector at time t becomes

$$\widehat{o_t} \quad A_c^{-1} o_t \quad A_c^{-1} b_c \tag{3.74}$$

This is essentially equivalent to transforming the feature vectors from a new speaker to lie in the acoustic space of the speaker independent models [Gales, 1998].

In constrained MLLR, closed form solutions do not exist for the computation of the transformation matrix. The transforms are therefore estimated iteratively.

3.3.1.3 Speaker adaptive training

Given a training corpus, in the normal acoustic model training procedure, the model parameters λ are estimated to maximise the likelihood of training data

$$\widehat{\lambda} \quad \arg\max_{\lambda} \quad L\ O;\lambda \tag{3.75}$$

In SAT paradigm [Anastasakos et al., 1996], the idea is to jointly estimate a set of speaker transforms and a set of canonical model parameters, such that the variations

due to speaker differences are captured in the speaker transforms ($G^{(r)}$ while the phonetic characteristics of the language are captured by the canonical models (λ_c). Canonical model parameters so estimated are normalised across speakers.

$$\widehat{\lambda_c}, \widehat{G} \quad \arg \max_{\lambda_c, G^{(r)}} \prod_{r=1}^{R} L\left(O^{(r)}; G^{(r)}, \lambda_c\right) \tag{3.76}$$

where, $\widehat{G} \quad \widehat{G}^{(1)}, \widehat{G}^{(2)}, \ldots, \widehat{G}^{(R)}$, and $O^{(r)}$ is the set of observation sequences associated with speaker *r*.

In practice, SAT is implemented in an iterative process as described below:

1. Train a speaker independent model using the Baum-Welch re-estimation process.
2. For each of the training set speakers estimate the CMLLR transforms using the last updated speaker independent model.
3. Normalise the feature vectors from the training set speakers using the corresponding transforms.
4. Retrain the SI model with the normalised feature vectors.
5. Repeat Steps 2-4 until the the likelihood scores stabilise for the training set. The final set of models are the canonical normalised models.

3.3.2 Maximum a posteriori adaptation

Maximum A Posteriori (MAP) adaptation provides a well defined mathematical framework for incorporating the prior information of the model parameters with the information provided by the adaptation data in the training process. The training process can be viewed as an interpolation between the original acoustic models and the maximum likelihood estimate of the adaptation data. While in maximum likelihood estimation the parameters λ are chosen such that the likelihood $p(x|\lambda)$ is maximised, in MAP estimation, parameters are set at the mode of the distribution $p(x|\lambda)\, g(\lambda)$ where $g(.)$ is the prior distribution of λ.

3.3.2.1 Standard MAP

The key issue in MAP adaptation is the choice of an appropriate prior distribution family. It is desirable to choose a prior density in the conjugate family which includes the kernel density $p(x|\lambda)$.

For a Gaussian mixture model with k mixture components, the kernel density $p\ x|\lambda$ for the data $x \quad x \quad _1 \ldots x_T$ is of the form

$$p\ x|\lambda \qquad \prod_{t \ 1}^{T} \sum_{k \ 1}^{K} w_k N\ x_t|\mu_k, \Sigma_k \tag{3.77}$$

The parameter set $\lambda \quad w \quad _1 \ldots w_K, \mu_1 \ldots \mu_K, \Sigma_1 \ldots \Sigma_K$ comprises the mixture weights w which form a multinomial distribution and the parameters of each component μ, Σ which are multivariate Gaussian densities. A sufficient statistic of a fixed dimension does not exist for λ and hence a conjugate prior cannot be readily specified.

This issue is addressed in the seminal work on MAP adaptation for continuous density HMMs by Gauvain and Lee [1994]. It is shown that by assuming independence between the weight parameters and the Gaussian density parameters, a prior distribution can be specified as a product of Dirichlet distribution and Normal Wishart distribution which are the conjugate pairs for multinomial and multivariate normal distributions respectively.

The update equations for the model parameters are derived using the Expectation Maximisation algorithm [Dempster et al., 1977].

Using MAP estimation, the update equations for a Gaussian component m in state j is given by

$$\mu_{jm} \quad \frac{\sum_t^T {}_1 \gamma_{jm}\ t}{\tau \quad \sum_t^T {}_1 \gamma_{jm}\ t} \bar{\mu}_{jm} \quad \frac{\tau}{\tau \quad \sum_t^T {}_1 \gamma_{jm}\ t} \mu_{jm} \tag{3.78}$$

where, $\bar{\mu}_{jm}$ is the maximum likelihood estimate of the mean of the adaptation data, μ_{jm} is the prior mean usually chosen from the previous iteration of the EM algorithm and τ is the hyperparameter that controls the bias between the prior information of model parameters and additional information from the adaptation data. τ is chosen heuristically depending on the strength of the prior models and the amount of amount of adaptation data available. Similar update equations exist for the other parameters.

Given a reasonable amount of adaptation data, MAP can be used to smooth or adapt the model parameters. An attractive property of MAP is its asymptotic convergence to maximum likelihood estimation as the amount of adaptation data increases. The main disadvantage of the use of MAP in its original form is that it updates only those parameters from their prior values which are observed in the adaptation data. In a large vocabulary system, there are typically thousands of Gaussians and this limits the usage of MAP when the adaptation data is very small. Various extensions to MAP have been proposed that aim to update the parameters associated with unseen data.

3.3.2.2 Structural MAP (SMAP)

Structural MAP [Shinoda and Lee, 1997, 2001] takes a slightly alternate view of adaptation. The core idea in this approach is to model the mismatch between the speaker independent mixture components and the adaptation data, and use such model as a prior to adapt the parameters in a maximum a posteriori sense.

Given an observation vector x_t, the mismatch of this observation with respect to every Gaussian mixture component is computed.

$$y_{mt} \quad \Sigma \ {}_{m}^{-1} \ {}^{2} x_t - \mu_m \tag{3.79}$$

If there were no mismatch, y_{mt} would be normally distributed with zero mean and unit variance $Y \quad N\ y;\ \bar{0}, I$. When there is a mismatch between the adaptation data and the SI models the distribution would be, $Y \quad N\ y; \nu, \eta$, $\nu \quad \bar{0}$ and $\eta \quad I$, where ν and η represent the shift and rotation needed to overcome the mismatch. This distribution is called the normalised pdf.

In SMAP procedure, all the mixture components in the speaker independent models are clustered into P clusters and normalised models are estimated for all the Gaussians in each cluster p in a maximum likelihood sense as shown in equations 3.80 and 3.81 respectively.

$$\tilde{\nu}^p \quad \frac{\sum_t^T \ {}_1 \sum_m^{M^p} \ {}_1 \gamma_{mt}^p y_{mt}^p}{\sum_t^T \ {}_1 \sum_m^{M^p} \ {}_1 \gamma_{mt}^p} \tag{3.80}$$

$$\tilde{\eta}^p \quad \frac{\sum_t^T \ {}_1 \sum_m^{M^p} \ {}_1 \gamma_{mt}^p \quad y_{mt}^p - \tilde{\nu}^p \quad y_{mt}^p - \tilde{\nu}^p}{\sum_t^T \ {}_1 \sum_m^{M^p} \ {}_1 \gamma_{mt}^p} \tag{3.81}$$

The adaptation step involves transforming the parameters of all the components in the cluster with its corresponding normalised pdf acting as the prior.

$$\tilde{\mu}_m^p \quad \bar{\mu} \ {}_m^p \quad \bar{\Sigma}_m^p \ {}^{1} \ {}^{2} \tilde{\nu}^p \tag{3.82}$$

$$\tilde{\Sigma}_m^p \quad \bar{\Sigma}_m^p \ {}^{1} \ {}^{2} \tilde{\eta}^p \left(\ \bar{\Sigma}_m^p \ {}^{1} \ {}^{2} \right) \tag{3.83}$$

To further improve the estimates, a hierarchical tree structure was proposed [Shinoda and Lee, 2001] to cluster the Gaussian components. With such a tree, the normalised model at a node is used as the prior distribution in estimating the normalised models for all the child nodes of that node. These MAP estimated normalised models at each node are then used to adapt all the Gaussian components associated with that node in maximum a posteriori sense.

3.3.2.3 Maximum a posteriori linear regression

While MAP adaptation provides a robust estimate for the parameters, its effectiveness is limited by the availability of sufficient adaptation data. The parameters of only the seen models in the adaptation data are re estimated. Linear regression based approaches on the other hand rely on the basis that several models are tied together. The adaptation data for all the tied models are pooled together to estimate the affine transforms for adaptation. Hence they give robust estimates even with small amounts of data. However with the availability of larger amounts of data, there is a tendency to overfit the parameters and the improvement in performance saturates quickly.

Maximum A Posteriori Linear Regression (MAPLR) [Chesta et al., 1999] was proposed as a solution that gets the best of both worlds. The idea is to effectively apply a global transform to all the models and further improve the estimates by local adaptation. The problem is posed as estimating the affine transforms ($W \quad A, b$) from the adaptation data using MAP criterion instead of ML criterion.

$$\widehat{W} \qquad \arg\max_{W} \quad P\ W|Y, \lambda \tag{3.84}$$

$$\widehat{W} \qquad \arg\max_{W} \quad P\ Y|W, \lambda\ P\ W \tag{3.85}$$

Under this formulation, it was suggested [Chesta et al., 1999; Siohan et al., 2001] that the prior distribution for the affine transforms be chosen as the matrix multivariate normal distribution

$$P\ W \quad |\Sigma|^{\quad - \quad n \quad 1} \quad |\Phi|^{-n \quad 2} exp\left\{ -\frac{1}{2} tr\ W - M \quad \Sigma^{-1}\ W - M\ \Phi^{\ -1} \right\} \tag{3.86}$$

where, M,Σ, and Φ are the hyperparameters for the distribution family, with M $\Re^{n \quad n \quad 1}, \Sigma \quad \Re^{n \quad n}, \Sigma \quad 0, \Phi \quad \Re^{\quad n \quad 1 \qquad n \quad 1}$ and $\Phi \quad 0$.

With the prior distribution assumed as above, the affine transform W can be estimated via the expectation maximisation formulation [Dempster et al., 1977].

With the availability of large amount of adaptation data, there arises a need to cluster the Gaussians and estimate an affine transform for each of these clusters. This leads to a requirement to define a large number of prior densities robustly. In Structural MAPLR (SMAPLR) [Siohan et al., 2002], a hierarchical prior structure is proposed as a tool to control the complexity of prior distribution estimation. Having defined a tree structure for the priors, the idea is to use the posterior distribution of W in a given

node as the prior distribution for the estimation of affine transform W_c for its child node. However, the posterior distribution of W at a given node does not belong to the class of matrix multivariate normal distributions. To resolve this problem, the true posterior distribution is approximated by a distribution from the matrix variate normal family having the same mode. This approximation is shown to work quite well and thus provides a good framework for robust adaptation.

3.3.3 Speaker space adaptation approaches

Under speaker space adaptation approaches, we discuss the Cluster adaptive training and Eigenvoices methods. These approaches are based on a slightly different paradigm compared with the previously discussed approaches. These methods look at the acoustic space as not only partitioned in terms of differences in phonetic characteristics but also by speaker characteristics.

3.3.3.1 Cluster adaptive training

In Cluster adaptive training (CAT) [Gales, 2000], the core idea is that speakers in the training set can be clustered by their acoustic properties. Hence each Gaussian component m is represented by a canonical model M

$$M_m \qquad \mu_m^1 \ldots \mu_m^C \tag{3.87}$$

where, the acoustic space of the Gaussian component is clustered into C classes, and μ_m^i represents the mean of the i^{th} class for the component m.

For a new test speaker (r) with a small amount of adaptation data, instead of hard assigning the speaker to a cluster class, the Gaussian components' means are chosen as a linear interpolation of all the classes either with or without a bias term

$$\mu_m^r \quad M \;_m \lambda_r \quad b \tag{3.88}$$

$$\mu_m^r \quad M \;_m \lambda_r \tag{3.89}$$

Here λ_r is the cluster weight vector $\lambda_r^1 \ldots \lambda_r^C$ for the speaker which needs to be estimated from the adaptation data in ML sense.

The key advantage of this approach is that for any new speaker, only a small number of cluster weight parameters need to be estimated. Hence rapid and robust adaptation becomes possible even with the availability of a small amount of adaptation data.

3.3.3.2 Eigenvoices

The underlying idea in the eigenvoices approach is that the acoustic characteristics of a speaker can be represented as a linear combination of the acoustic characteristics of other speakers. In eigenvoices [Kuhn et al., 2000], from a set of speaker dependent models and the speaker independent model, high dimensional supervectors (dimensionality D) are created by concatenating the parameters that need to be adapted. The order of the parameters in the supervectors is the same for all the speakers.

From these R ($R \quad D$) supervectors, using dimensionality reduction techniques such as principal component analysis, R eigenvectors of dimension D are determined. These are then ordered in terms of their eigenvalues and only the top K along with the supervector of speaker independent model ($e\ 0$) are retained as the basis vectors in the eigenspace ($K < R$).

Each new speaker is projected as a point P in this space as shown in equation 3.90

$$P \quad e\,0 \quad w\,1\,e\,1 \quad \ldots \quad w\,K\,e\,K \tag{3.90}$$

The projection weights for a speaker are estimated from the adaptation data using maximum likelihood eigendecomposition (MLED) [Kuhn et al., 2000]. The standard ML approach to maximise the auxiliary function results in K equations with K unknown weight parameters which can solved using Gauss-Jordan method.

Similar to CAT, the number of parameters to be estimated for a test speaker are small and hence can be robustly estimated even with small amount of adaptation data.

3.4 Automatic age recognition

Automatic estimation of speaker age has been a growing topic of research over the past few years. An age estimate of a speaker can be used as a useful cue to adjust spoken dialogue system's behavior to suit the communication behavior of the speaker age group [Shafran et al., 2003]. It could also be used to give appropriate information, and in the mechanism of information delivery. However estimation of age based on voice has not been an easy task to solve.

Though speech production deteriorates with age, there is no clear correlation between chronological age and speech. Some voices tend to show signs of ageing in 40 years while the voices of some healthy elderly people above the age of 65 years do not show signs of ageing. The profession of speakers (in particular whether the profession

involves lot of speaking/shouting), smoking habits, and health conditions seem to have a bearing on the vocal ageing.

Most of the automatic age estimation techniques hence try to achieve performance comparable to perceived age. Perceived age is the age of a voice as perceived by human listeners. Several approaches using cepstral features and perturbations in voice have been used to classify speakers in age groups and compare the results with subjective listening tests.

Cepstral features [Davis and Mermelstein, 1980] have been investigated by many researchers for discrimination of age. The use of these features has been motivated by the fact that they are robust for speech and speaker recognition applications. Minematsu et al. [2002] used MFCC and Δ MFCC coefficients as acoustic features to create Gaussian Mixture Models (GMM) of elderly speakers and non-elderly speakers. All the speakers (in training and test sets) were classified as elderly or non-elderly based on listening test. The test identified perceived elderly speakers with 90.9% accuracy. Speech rate and local power perturbations were then added as additional features and the experiment repeated. The additional features gave a better discrimination of age at 95.3% identification rate. Shafran et al. [2003] used a HMM classifier based on cepstral features and pitch to recognise age, gender, dialect and emotion. Results on age classification showed 68.4% correct classification using only cepstral features which increased to 70.2% using both cepstral features and F_0 features. In another approach, Ajmera and Burkhardt [2008] reported that the slow moving temporal envelope of the Cepstral features called Modulation Cepstrum coeffients give better age discrimination than with the use of cepstral features themselves.

Müller et al. [2003] used various measures of Jitter and Shimmer for classification of speakers to their age group. Jitter and shimmer as acoustic measures provided consistently good results in age discrimination for a range of machine learning approaches for classification including Artificial Neural Networks, K Nearest Neighbours, Naive Bayes, and Support Vector Machines. Müller [2006] extended the acoustic features from jitter and shimmer to also include harmonics to noise ratio, articulation rate, and frequency and duration of speech pauses and the increased the complexity of classification problem to 8 classes representing children, teenagers, adults and seniors for both the genders. 63.5% correct classification accuracies were achieved in this task. The system was further improved [Müller and Burkhardt, 2007] by using a classifier combining GMMs using frame based Mel Frequency Coefficients and Support Vector Machines using long term pitch features. The system achieved classification perfor-

mance close to human listeners.

In a recent study [Metze et al., 2007], a comparative study of four different approaches to age and gender classification has been made. The challenge was to classify a speaker into one of the groups of children, young males, young females, adult males, adult females, senior males and senior females. The four systems used were A) A parallel phoneme recogniser with separate models for each class B) A classifier using dynamic Bayesian networks to combine various prosodic features C) A system based on linear prediction analysis and D) GMMs based on MFCC features. Parallel phoneme recogniser seems to give the best performance for longer utterances but its performance drops with shorter utterances. The system using prosodic features fared well for both short and long utterances.

More recently, an age recognition challenge to classify speakers into one of the 4 subgroups (Child, Young, Adult and Senior) was organised as a subchallenge in the paralinguistic challenge at interspeech [Schuller et al., 2010]. The system that achieved the best accuracies [Kockmann et al., 2010] is based on a fusion of GMM-UBM models, discriminatively trained models, SVMs, eigenvoice and anchor based models. The accuracies obtained in age recognition were around 56%.

From all these studies, it is clear that there is no clear correlation between the chronological age and the vocal age. However various prosodic features are good indicators of ageing voice, though they may not be used to predict the actual age accurately.

3.5 Automatic speech recognition on older voices

Studies on various changes in the characteristics of speech signal observed in older voices have been discussed in 2.3. It was also seen from the review in 3.4, that such changes in voice also leave acoustic cues which can be exploited by human listeners as well as machines to infer the speakers age with reasonable accuracies.

As discussed in 2.3, less precise articulation is often associated with older voices. Shuey [1989] conducted an interesting experiment in order to understand the speech intelligibility differences between younger and older voices. Speakers from the two age groups were asked to read CVCs embedded in a carrier phrase and the listeners had to transcribe the target word. It was found that significantly higher number of errors were made for the older speakers' utterances. In an experiment to understand socio psychological meaning of older people's language and communication, Williams and Giles [1992] report that older people's voices were rated perceptually lower than

younger voices by a set of young listeners and the recall of the message of the older people was also found to be significantly lower than those of younger voices.

While there have been numerous studies on the effects of ageing on the voice, there has been limited work to understand the performance of ASR systems on ageing voices.

In an experiment to understand whether special acoustic models are required for automatic speech recognition of elderly voices [Baba et al., 2004], it has been observed that the recognition accuracies of elderly voices above 70 years of age are 9-12% lower than adult speakers using speaker independent acoustic models. While the drop in performance for aged females was around 4-7%, it was significantly higher at 16-18% for aged males. A relative increase of 5-8% in accuracies was achieved when acoustic models trained using elderly speech was used. It was also observed that the acoustic models trained on elderly voices served as better baseline models for speaker adaptation than those trained on younger adult voices. These results are consistent with the observations made by Anderson et al. [1999] in which acoustic models trained with elderly speech gave 12% better accuracies with elderly voices as compared to those with acoustic models trained on non-elderly speech. Elderly men were found to have substantially higher WERs. Further improvements in accuracies could be achieved using gender and age group specific models and such results were found to be comparable to speaker adapted models with VTLN normalised features.

In a study of speech recognition for children and the elderly [Wilpon and Jacobsen, 1996], it has been found that the error rates increase dramatically for voices below 15 years and above 70 years of age. They also observe that while accuracies could be improved for younger voices by modifying the front end of the speech recogniser and with additional training data, such improvement in results could not be replicated for older voices.

These research results indicate that there are differences in the acoustic properties of younger and older adults and lay the motivation for the work presented in this thesis. It is of interest to investigate the possible causes for these differences in WERs and to explore some possible ways to improve ASR accuracies for older voices.

Chapter 4

ASR accuracy on ageing voices: Baseline Experiments

As discussed in section 3.5, ASR error rates have been reported to be higher for older adults compared to younger adults. In this chapter, the ASR accuracies for younger and older adults are presented and compared on three different corpora. The three corpora are first described in detail. The experimental setup for each of these corpora along with the baseline results are discussed in the subsequent subsections.

4.1 Corpora

One of the main challenges in working with older voices is the lack of speech corpora for this domain. Most of the speech corpora used in ASR research are collected from younger and middle aged adult speakers. The following three corpora collected in three different continents have been used in this research work

SCOTUS Corpus - US English

MATCH corpus - UK English

JNAS Corpus - Japanese

All these three corpora have a good representation of speech from older speakers. They are also reasonably balanced in terms of gender (male and female) and age (younger and older) of the speakers. Each of these corpora also captures a distinct speaking style. SCOTUS is a rehearsed spontaneous speech while the MATCH corpus captures typical human interaction style with spoken dialogue systems in the form of

short and dialogue driven utterances. JNAS contains several hours of read speech of newspaper articles.

The three corpora and described below in detail and the advantages and disadvantages of the usage of each of these corpora for this research are also discussed.

4.1.1 SCOTUS

The SCOTUS speech corpus is a collection of the audio recordings of the oral arguments of the Supreme Court of The United States. These recordings have been made public under the Oyez project [3]. Each recording's duration is about one hour and consists of speech from the advocates and judges arguing the case. These recordings were archived on reel-to-reel tapes, which were later digitised and made public.

Although the recordings from the 1980s to the present date are currently available online, complete transcripts with speaker information are available only from the later half of 1990s. Hence only those audio files annotated with speaker tags were used in our experiments. In all, the experimental corpus contains 534 recordings. It consists of speech from 10 judges over several years and speech from about 500 advocates. The birth dates of the judges are known and hence their age at the time of an argument can be precisely calculated. The birth dates of the advocates are not easily available, hence wherever the dates were not available, their age has been approximated by using the year of their law graduation and assuming their age at graduation to be 25.

The corpus available on the Oyez website is not readily usable for ASR experiments. Each of the audio files is about 1 hour in duration with several speaker turns and the transcripts have digits to represent years and case numbers etc. Several preprocessing steps were involved as detailed below to make the corpus usable for ASR experiments.

1. The first step involved text normalisation. The punctuations and speaker turn tags were first removed from each audio transcript to get a plain text transcript of the entire audio. The speaker-sentence correspondence was stored in a separate metafile. The text contained digits in several forms viz., years, supreme court case ids, currency and normal numbers. Context based rules were setup to convert the digits to text. This process involved several iterations and tweaking the rules to get most conversions correct. The text normalisation was manually checked with random sampling.

[3] `http://www.oyez.org`

2. The audio files available for download are in MPEG format. These were converted to 16KHz 16 bit waveforms.

3. In order to obtain the sentence boundaries and speaker turn alignments, each of the 1 hour long files was force aligned using acoustic models trained on 73 hours of meetings data recorded by the International Computer Science Institute (ICSI), 13 hours of meeting corpora from the National Institute of Standards and Technology (NIST) and 10 hours of corpora from the Interactive Systems Lab (ISL) [Hain et al., 2005a]. These models will be referred to as ICSI-NIST-ISL models in this thesis.

4. Each utterance was renamed as shown below to reflect all the available meta data. *caseId_SpeakerId_Age_Sex_StartTime_EndTime*

4.1.1.1 Advantages and Limitations

One of the advantages of this corpus for ASR experiments is that the recording setup for the Court proceedings has remained the same over a period of time and hence the variations in noise and microphone characteristics are minimal. This reduces the confounding effect of recording and channel conditions on ASR WERs.

The language used in the Supreme Court is formal and is fairly similar across all the speakers. This allows us to assume minimal variability in the language models and focus on the acoustic models.

The other advantage is that the data from the supreme court judges is available over several years and thus allows longitudinal analysis.

One of the limitations of this corpus in that the number of older speakers (above 60 years) of age is quite limited. The corpus is also skewed by gender with most of the corpus being from male speakers.

4.1.2 MATCH

The MATCH corpus [Georgila et al., 2009] was recorded at the University of Edinburgh for a cognitive psychology experiment that investigated the accommodation of cognitive ageing in spoken dialogue interfaces [Wolters et al., 2009]. Speech utterances were recorded from 24 younger users (aged 18-29 years, mean 22) and 26 older users (aged 52-84 years, mean 66) in a wizard of oz (WOZ) system where each user

booked health care appointments using nine different simulated spoken dialogue interfaces. A total of 447 dialogues were recorded using an EDIROL R01 digital recorder and a sampling frequency of 44.1 kHz. The dialogues contain 3.5 hours of speech in total. All dialogues were transcribed orthographically by a trained annotator using the tool 'Transcriber' [Barras et al., 2001] in accordance with the AMI transcription guidelines, which were used for creating the AMI meetings corpus [Carletta, 2007].

The corpus has been annotated semi-automatically with dialogue acts and information state update information [Georgila et al., 2008]. Overall, the speech corpus contains 1680 speech spurts [4] from older adults and 1369 speech spurts from younger adults.

4.1.2.1 Advantages and Limitations

Since the MATCH corpus was created for a cognitive psychology experiment, dialogue structure, appointment scenarios and system vocabulary are tightly controlled. As a result, the vocabulary is much less diverse and the language is more formulaic than that of corpora which are recorded for speech technology research. It is also relatively small compared to other speech research corpora. Despite these disadvantages, the MATCH corpus is one of very few corpora that has a good balance of older and younger speakers. It contains highly detailed dialogue act and information state annotations.

4.1.3 JNAS

The Japanese corpora used in our experiments consists of 2 sets namely JNAS (Japanese Newspaper Article Sentences) and S-JNAS (Senior-JNAS). Both these corpora are read speech of sentences from Mainichi newspaper article sentences and a set of phonetically balanced (PB) sentences from the Advanced Telecommunications Research (ATR) institute. The JNAS corpus is predominantly from young and middle aged adults. The JNAS corpus was collected by the Acoustic Society of Japan (ASJ) [Itou et al., 1998, 1999]. The S-JNAS corpus is comprised of speech from older adults in the age group of 60-91 years. This corpus was collected at the Nara Institute of Science and Technology (NAIST) aimed at development of speech recognition technologies for older people. Although these corpora were collected separately at different points in time (S-JNAS was collected about 6 years after JNAS), they were recorded under

[4] Here we refer one continuous segment of speech spoken by a user in his/her interaction turn with the spoken dialogue system as a speech spurt

similar recording conditions in booths using Sennheiser head mounted microphones at 16KHz sampling rate as 16 bit waveforms. This is particularly useful since noise and channel conditions are similar across the two corpora in addition to the sentences spoken being the same. The main factor that varies between the two corpora is the age range of the speakers as seen in Fig. 4.1.

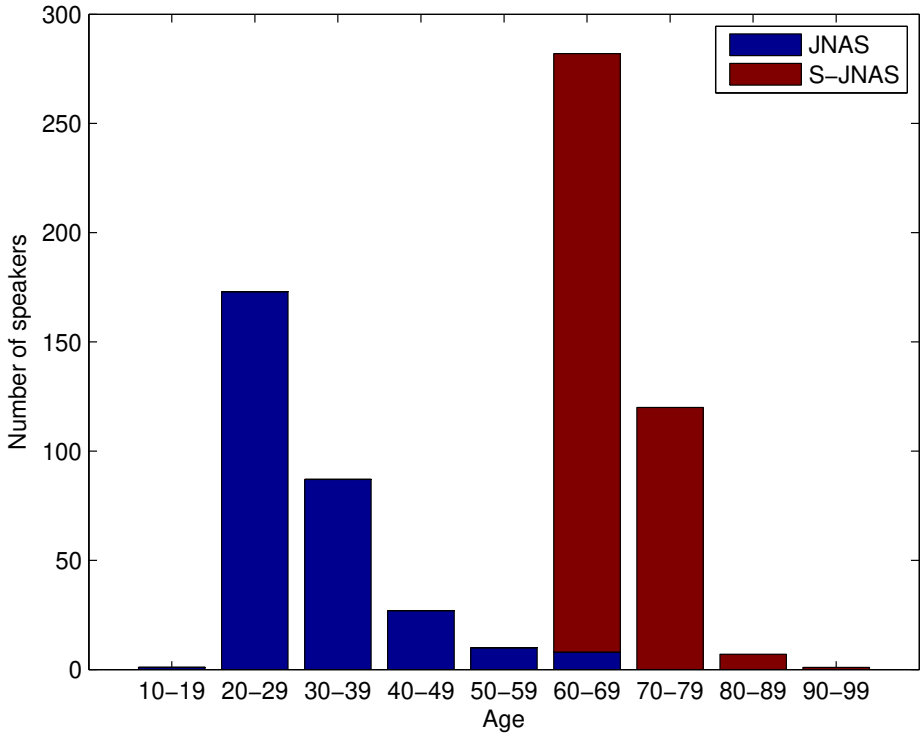


Figure 4.1: Age distribution of speakers in the JNAS and the S-JNAS corpora

More specific details of the corpora are listed below.

JNAS: It comprises read speech from 306 speakers (153 Male and 153 Female). There are in all 155 text sets, each set consisting of 100 sentences from newspaper articles and 50 ATR-PB sentences. Each set is usually read by one male and one female speaker. Thus each speaker has around 150 utterances making a total of about 45000 utterances in all.

S-JNAS: The S-JNAS corpus has a predefined training and test set of speakers. The training set has 301 speakers (151 Male and 150 Female) and the test set has 101 speakers (51 Male and 50 Female). Each speaker's recordings include one set of 100 sentences from newspaper articles and two sets of 50 sentences each from ATR-PB sentences giving a total of 200 utterances. The total amount

of speech data is about 133 hours for the training speakers and about 44 hours for the test speakers.

4.1.3.1 Advantages and Limitations

The main advantage of using the JNAS corpus in this research work is the availability of a large amount of speech data from a large number of speakers. The corpus is well balanced in terms of speaker age groups and gender. The utterances spoken by the speakers in the JNAS and S-JNAS corpus are the same. This allows us to design the experiments such that the test sets for younger and older adults have the same speech utterances. This nullifies the impact of language models in the comparative results.

4.2 ASR WERs on older voices

In this section, the ASR experimental setup using the three corpora discussed above are described. State of the art ASR systems were built using Hidden Markov Model toolkit (HTK)[5] [Young et al., 2006] and the ASR WERs of older and younger voices are compared.

4.2.1 Experiments with SCOTUS corpus

4.2.1.1 Comparison of ASR WERs on younger adult and older adult voices

Feature extraction

The SCOTUS corpus in MP3 format was first converted to 16KHz wav format and then parametrised using perceptual linear prediction (PLP) Cepstral features. A window size of 25ms and frame shift of 10ms were used for feature extraction. Energy along with 1st and 2nd order derivatives were appended giving a 39-dimensional feature vector.

Cepstral means and variances were computed for each speaker in each recording. These were then used to normalise the feature vectors to minimise any channel induced affects.

Acoustic models

The acoustic models were trained on 90 hours of speech data from the advocates. A significant portion of the entire corpus is from males, hence the training data set is also

[5]HTK version 3.4 `http://htk.eng.cam.ac.uk`

similarly skewed in favour of males with around 77 hours of speech from males and 13 hours of speech from females. As mentioned in section 4.1.1, the ages of some of the speakers used in the training set are unknown. The distribution of the ages of speakers (where known) is shown in Figure 4.2. Although it does not contain all the data, it is suggestive of the fact that that the training set speakers are predominantly younger adults.

The acoustic models have been trained as cross-word context-dependent triphone HMMs, each state modeled as 18 component GMM for all speech phonemes and 36 component GMM for non-speech (sil & short pause) models respectively.

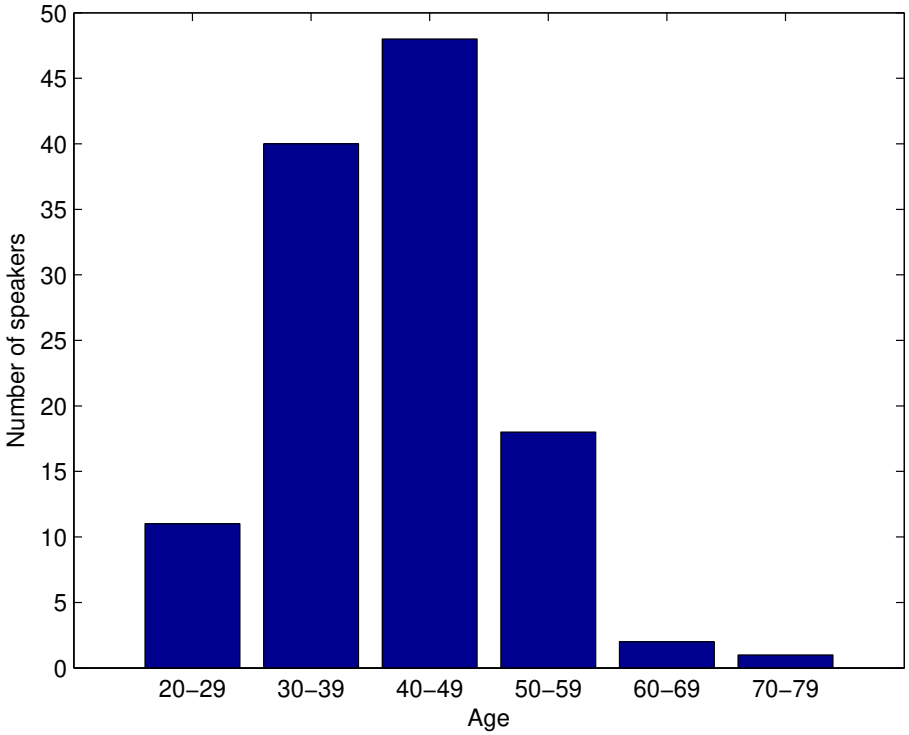


Figure 4.2: Age distribution of speakers in the training set of the SCOTUS corpus

Language models

The language models were constructed from the transcripts of 260 United States Supreme Court recordings from the 1970s. Back-off bi-gram language models were constructed from this data. The vocabulary consists of 23445 words. The pronunciations in the AMI lexicon were used for those vocabulary words common to AMI vocabulary [Hain et al., 2005a] and the pronunciations for the rest of the vocabulary words were generated using the Festival speech synthesis system [Taylor et al., 1998].

Test utterances

For the *younger adult* test set, speech utterances from 27 speakers (23 Male and 4 Female) in an age range of 30-45 were chosen. For the *older adult* test set, speech data from 12 speakers (10 Male and 2 Female) in the age range 60-85 were used. The test speaker set is disjoint from the training set speakers. 10 utterances from each test speaker were kept aside for speaker adaptation and the remaining utterances formed the test set. In all, the *younger adult* test set comprises of 4964 utterances (14.5 hours) and the *older adult* test set comprises of 6652 utterances (19 hours). The perplexity of the language model on the two test sets and the OOV rates are shown in Table 4.1.

Language Model Perplexity and OOV rate (%)		
	***Younger adult* test set**	***Older adult* test set**
Perplexity	178.3	169.7
OOV Rate	3.8%	4.3%

Table 4.1: Perplexity and OOV rate for the *younger adult* and *older adult* test sets in SCOTUS corpus

Baseline results

The ASR word error rates on *younger adult* and *older adult* test sets are shown in Figure 4.3. The results show a significant difference of 9.3% absolute higher WERs for older voices as compared to younger voices. The WERs difference for males is 8.2% absolute while for females it is 13.3%. The differences in WERs are statistically significant with $p < 0.001$ using the Mann-Whitney test [Mann and Whitney, 1947]. A possible reason for such high WERs for female speakers is the inadequate representation of females in the training set.

Standard MLLR mean adaptation was used to see if speaker adaptation could alleviate age induced errors in ASR. Using the adaptation set of 10 utterances for each speaker, MLLR transforms were computed for each speaker and used in decoding the test utterances. The difference in WERs even with MLLR speaker adaptation is 9.1% absolute.

One of the main sources of inter-speaker variability in acoustic features is the variation in vocal tract dimensions. Vocal Tract Length Normalisation (VTLN) is a standard approach used to overcome this variability. Vocal tract length normalised acoustic

models were constructed using an iterative approach as described in section 3.2.2. Using the normalised models, warping factors were estimated for each of the test speakers from the adaptation set utterances. With VTLN, the improvements in WERs for younger adults are higher than those for older adults leading to a difference in WER of 9.9% absolute between the two age groups.

Figure 4.3 also shows the comparative results with Speaker adaptive training. Using an iterative process as described in section 3.3.1.3, canonical models were trained. For each test speaker, using the same adaptation set as used above, CMLLR transforms were computed with respect to the speaker normalised canonical models. SAT gives the best improvements in WERs. However, the difference in WERs between the two age groups is still 9.8% absolute.

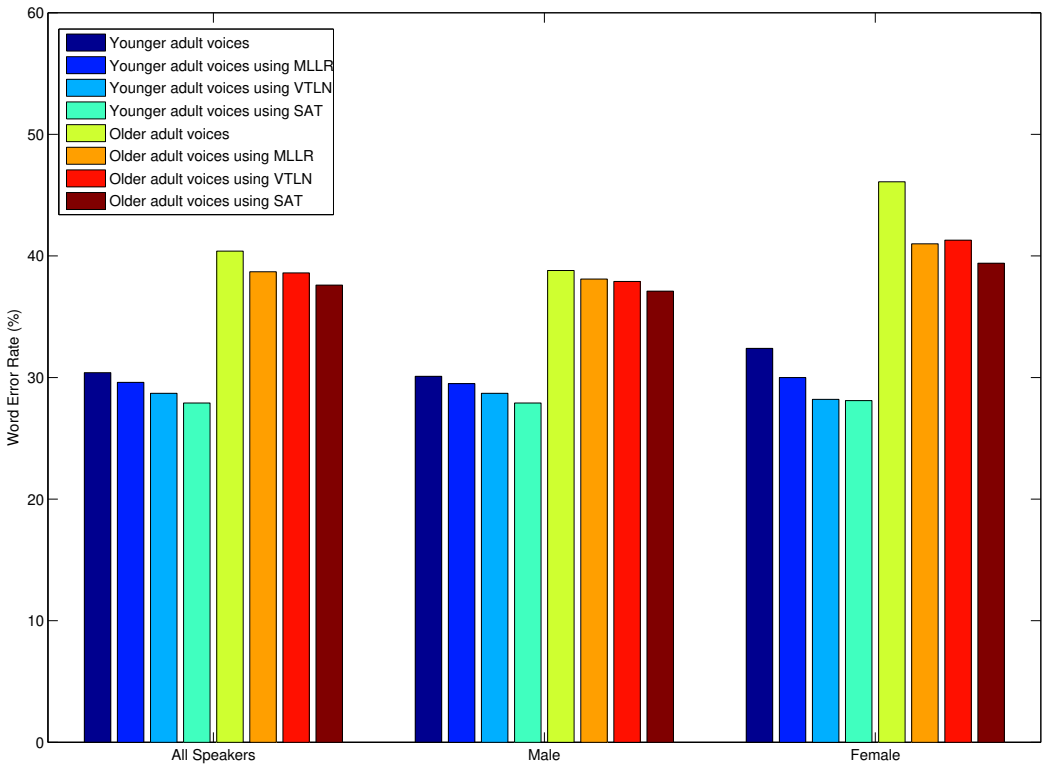


Figure 4.3: Comparison of WERs on *younger adult* and *older adult* voices in the SCOTUS corpus. Refer tables: A.1, A.2, A.3 and A.4

From the results, we observe that though speaker adaptation and speaker normalisation improve the recognition accuracies, the gap between the WERs for adult and older voices is not bridged. The results for females may not be a true representation

of the difference as the sample set is very small, but overall the difference in WERs seems to be large enough for investigation into the possible causes.

4.2.1.2 Longitudinal Study of ASR accuracies on older voices

Speech data of the Supreme court judges is available over a period of several years. To understand how the ASR accuracies vary for an older speaker with age, a longitudinal ASR experiment was setup. 200 utterances from each year for each speaker from the 7 judges (5 Male and 2 Female) was used as test set. Speakers with IDs $02, 03, 04, 05$ and 08 are males while speakers 07 and 10 are females. The results have been plotted in Figure 4.4. It can be seen that though there are small fluctuations in WER over each year, there is a general pattern in the results showing increase in WER gradually as the age increases for some speakers.

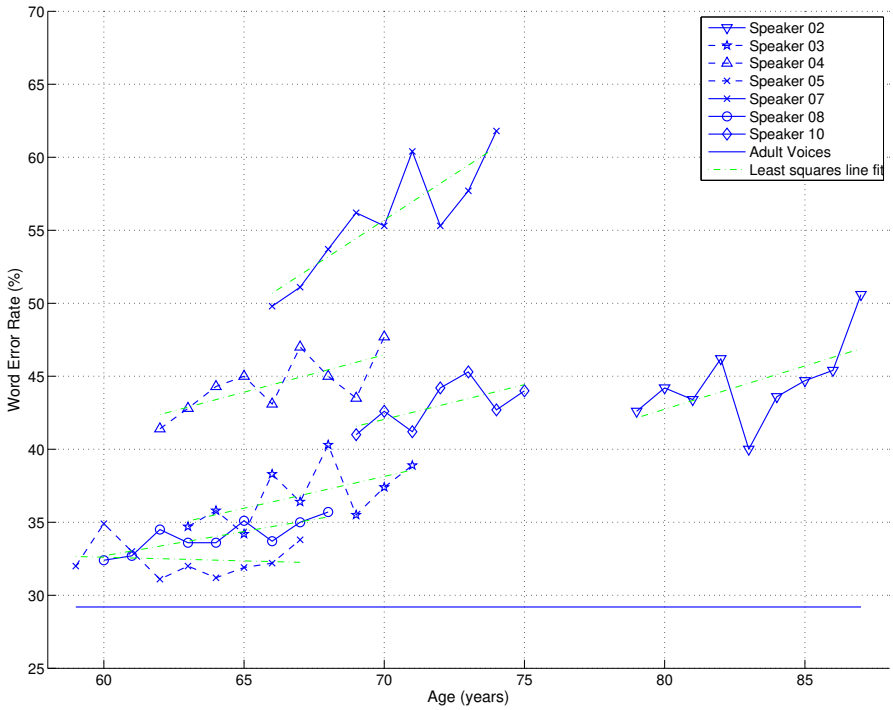


Figure 4.4: WERs (%) with increasing age on *older adult* voices in the SCOTUS corpus.

For each speaker, MLLR transform matrices were computed for mean adaptation. The regression class tree consisted of 2 classes, one for speech and one for non-speech. The longitudinal WER plots for all the speakers with speaker adaptation are shown in

Figure 4.5. A least square line fit for the WER over several years for each speaker are also plotted to understand the longitudinal trend.

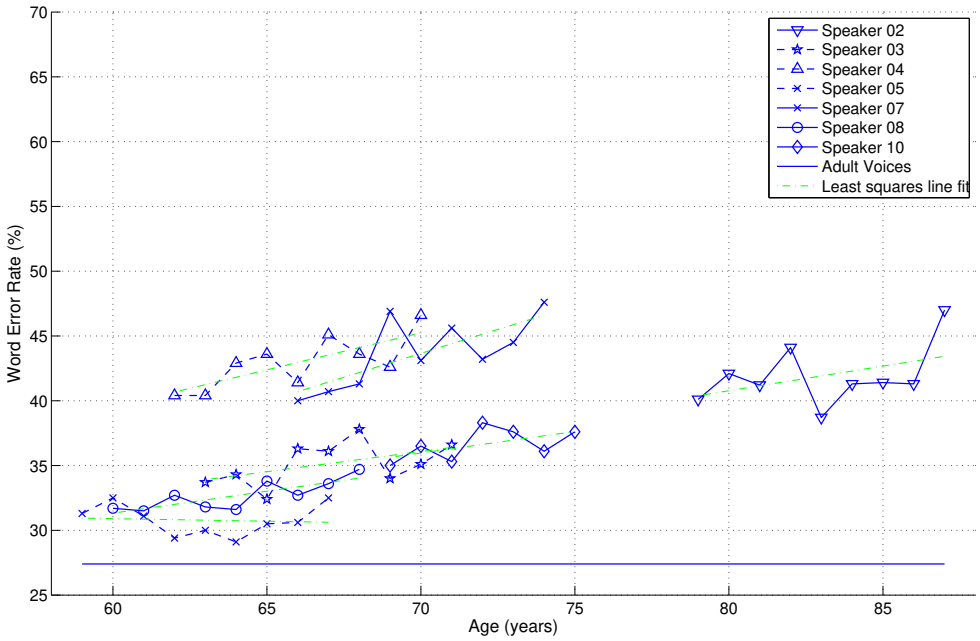


Figure 4.5: WERs (%) with increasing age on *older adult* voices in the SCOTUS corpus with speaker adaptation

The longitudinal study results indicate that the WER gradually increases with age during old age. Since the number of utterances for each year for each speaker was limited, variations in the WER are expected, however the least squares line fit for WER of all the speakers have a positive slope which suggests an increase in WER with age especially after 65 years of age. For speakers $04, 07$ and 08, F-tests show that there is strong evidence ($p < 0.01$) of a linear trend, and we conclude that in these cases WER is indeed increasing with age.

From the regression plots, it is seen that the increase in WER with age varies across speakers and the rate of increase differs for two speakers at the same age. This suggests that there is no clear correlation between chronological ageing and vocal ageing across speakers.

Longitudinal studies on elderly voices using MLLR adaptation also show a gradual increase in WER with age. For the case where MLLR has been used, we find that only speakers 04 and 08 show statistically significant evidence of a linear trend of increasing WER with age. The slopes of the longitudinal plots of each speaker using MLLR

adaptation are less than those without adaptation, indicating that speaker adaptation can reduce the age related impact to some extent.

4.2.2 Experiments with MATCH corpus

As seen from the results on SCOTUS corpus, the WERs on older voices are significantly higher than those of younger voices. We perform similar baseline experiments on the MATCH corpus. In this set of experiments, we examined the affect of age-specific language models and acoustic models on speech recognition accuracies. Since the amount of data available for each speaker is quite limited in this corpus, a 'leave one out' strategy has been used in the experimental design.

4.2.2.1 Impact of language modeling

Design

The aim of this experiment was to assess the effect of the differences in interaction style between younger and older users in the MATCH corpus, on the language modeling component of the speech recogniser and consequently on ASR accuracies. From the transcripts of the MATCH corpus, the following bi-gram language models were constructed:

1. from all the utterances of the older speakers *(LM-Older)*
2. from all the utterances of the young speakers *(LM-Young)*
3. for each test speaker, from the entire corpus excluding the data from the test speaker *(LM-All-1)*
4. for each older test speaker, from the corpus of all the older speakers excluding the data from the test speaker *(LM-Older-1)*
5. for each young test speaker, from the corpus of all the young speakers excluding the data from the test speaker *(LM-Young-1)*

Since the amount of data in the MATCH corpus is not sufficient to build acoustic models from scratch, we used the speech from other corpora for this purpose. ICSI-NIST-ISL acoustic models described before were MAP adapted with 13 hours of speech from 32 UK speakers from the Augmented Multiparty Interaction (AMI) corpus.

For each older test speaker, three ASR experiments were performed, keeping the acoustic model fixed and using different language models for the speaker viz., *LM-All-1*, *LM-Older-1* and *LM-Young*. Similarly, ASR experiments were repeated for each of the young speakers using the language models: *LM-All-1*, *LM-Young-1* and *LM-Older*.

Results

Goodness of fit of the language model on a test set was measured using perplexity. We also assessed the number of OOV words. We found that language models trained on younger users were a bad fit to the language of older users, whereas data from the older users allowed us to model the language patterns of younger users reasonably well. In particular, models trained on younger users only did not contain many of the words older people used. Detailed results are shown in Table 4.2.

Perplexity and OOV rate (%)			
Test set	**Language Model**	**Perplexity**	**OOV Rate (%)**
Younger	*LM-Older*	5.44	1.38
Older	*LM-Young*	19.18	15.57

Table 4.2: Comparison of the perplexities of the language model and OOV rates to understand the goodness of fit of the language models trained on younger users data for older users test set and vice-versa on MATCH corpus

Figure 4.6 shows ASR WERs using different language models as explained above, averaged over all the young speakers and older speakers respectively. As we would expect from the results presented in Table 4.2, we find that WERs for older speakers are particularly high when using the language models of the younger speakers. This is due to the mismatch between the older and younger users' interaction styles. Clearly, we need age-appropriate data to build adequate language models for older speakers.

4.2.2.2 Impact of Acoustic modeling

Design

In this set of experiments, we examined the impact of differences in the acoustics of older and young speakers on ASR accuracies. In order to isolate the effect of the acoustic models, we only used the language model *LM-All*, which contains all utterances in

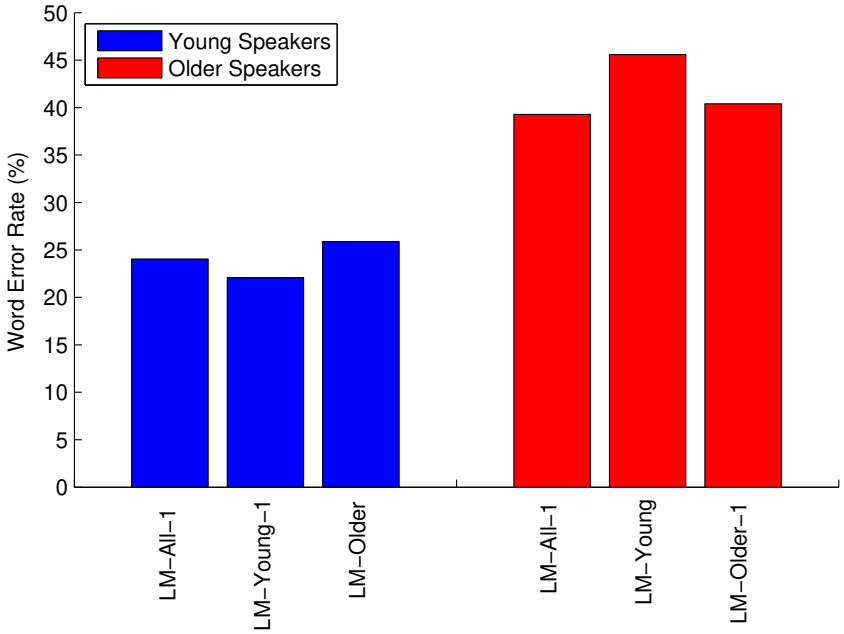


Figure 4.6: WERs (%) for young and older speakers of the MATCH corpus using different language models. (Refer Table A.7)

the MATCH corpus, for this set of experiments. The acoustic models described in the previous experiment (ICSI-NIST-ISL models adapted with AMI data) were used as the baseline models. For each of the old speakers, two acoustic models were created by maximum a posteriori adaptation of the baseline models using the speech from either the rest of the old speakers excluding the test speaker (*AMI + MATCH older-1*) or speech from the young speakers (*AMI + MATCH younger*). Acoustic models were similarly created for each young speaker with the speech data from all the older speakers (*AMI + MATCH older*) and the speech data from the rest of the young speakers (*AMI + MATCH young-1*).

Results

Figure 4.7 shows average WERs for both the young speakers and the older speakers. We observe that the WERs for older speakers are higher than those for younger speakers by 11% absolute using the baseline acoustic models. Adapting the models with speech from a new domain (i.e. appointment scheduling) is expected to reduce

the WERs for the test data in the new domain. While adapting the baseline models with older speakers from the MATCH corpus (AMI + MATCH older) brings down the WERs for young speakers, the results are even better with adaptation using speech from other younger speakers in the same corpus (AMI + MATCH young-1). The results for older speakers in Figure 4.7 are quite interesting, Contrary to the belief that speech from a new domain should help in creating better models for the new domain, adapting the baseline models with speech from the younger speakers of MATCH corpus (AMI + MATCH young) deteriorates the performance for the older speakers in the same corpus. Hence, there is a clear mismatch in the acoustics of older and young speakers resulting in a higher WER for older speakers.

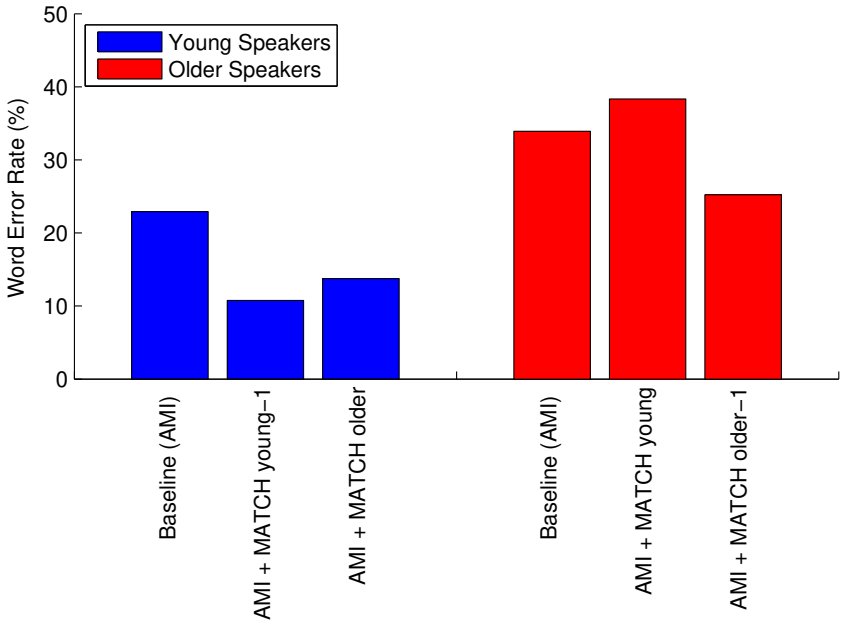


Figure 4.7: WERs (%) for young and older speakers of the MATCH corpus using different acoustic models. (Refer Table A.8)

4.2.3 Experiments with JNAS corpus

Most of the components from the Japanese ASR Toolkit [Kawahara et al., 1999] were used in our setup.

4.2.3.1 Acoustic models

Mel frequency cepstral coefficients with 12 filter banks were computed every 10msec. Appending delta and Energy coefficients, the feature vectors had a dimension of 25. Cepstral mean was subtracted for each utterance.

The acoustic models were trained using HTK [Young et al., 2006] as continuous density HMMs. The phoneme set comprised of 43 phonemes as defined by the Acoustic Society of Japan. 3 Pause models *SilB*, *SilE* and *sp* for beginning and end of utterance and short pause between words respectively are used in the phoneme inventory. Context dependent triphone HMMs with 5 states per model and each state modeled as 16 component GMM were trained.

4.2.3.2 Lexicon

Japanese texts are written without spacing between them. Using a state-of-the-art Japanese morphological analyser named ChaSen [Matsumoto et al., 1999] developed at NAIST, the text from Mainichi newspaper was segmented into word chunks (morphs). After the segmentation, the next step in text processing is conversion from Kanji (chinese characters) to Kana conversion which is equivalent to a Grapheme to Phoneme mapping. The Kana transcripts are further converted from orthographic to phonemic katakana form. A vocabulary of 20K [Kawahara et al., 1999] constructed from the most frequently used words (morphemes) in the Mainichi newspaper was used in our experiments.

4.2.3.3 Language models

Word 2-gram and 3-gram language model with back off smoothing constructed from 65 million words (morphs) in the Mainichi newspaper were used in the decoding.

4.2.3.4 Decoder

The open source large vocabulary speech decoder Julius [Lee et al., 1998] was used in decoding. It uses a forward backward two pass algorithm. In the first pass a frame synchronous beam search in the forward direction outputs a word lattice. In the second pass the lattice is re-scored in the reverse direction using stack decoding approach.

4.2.3.5 Evaluation setup

Training Set

The training set comprises of 205 speakers (about 31638 utterances) from JNAS corpus and 187 speakers (28332 utterances) from the SJNAS corpus. Thus the training set is balanced in terms of age and gender.

Test Set

The test set for SJNAS corpus was predefined with 101 speakers. An equal number of test speakers were chosen from the JNAS corpus such that the utterances from the younger and older adults corresponded to the same sentences. This factors out the differences in the ASR accuracies between speakers of the two age groups due to language use pattern. The differences in accuracies if any, would be purely due to the differences in acoustic characteristics between speaker sets.

The test set comprises of utterances from 101 older adults (5024 utterances, approx 50 utterances/speaker) and 101 younger adults (4099 utterances, approx 50 utterances/speaker).

4.2.3.6 Baseline Results

Table 4.3 shows the baseline comparative accuracies for the two age groups using SI acoustic models. The difference in WERs between younger and older adults is found to be 4.5% absolute. While the WER increases by 1.4% in older females as compared to younger females, the difference is more prominent in male speakers with a difference of 7.7% absolute.

	Younger Adults	**Older Adults**
All speakers	15.9	20.4
Male	16.2	23.9
Female	15.5	16.9

Table 4.3: Comparison of WERs (%) of younger and older adults in the JNAS corpus

The split up of the WERs for the older speakers in age groups of 60-69 and 70-79 are shown in Table. 4.4. The WERs for older males is particularly high in the age group of 70-79. The WERs for older females in 70-79 are quite low. Since there are only six female speakers in this subset, the result may be somewhat biased.

Age Group	#Speakers	Older Male Adults	Older Female Adults
60-69	80 (M:36 F:44)	21.3	17.0
70-79	21 (M:15 F:6)	30.0	14.2

Table 4.4: WERs (%) for older adults in different age groups in the JNAS corpus

4.3 Summary

The WER for older voices was found to be significantly higher than for younger voices from the baseline experiments on the three corpora. Use of standard speaker normalisation and speaker adaptation approaches improve the performance for older speakers marginally. However, the difference in the WERs for the two age groups persists. The results on the MATCH corpus also highlight the fact that the interaction style of older people with spoken dialogue systems is significantly different from younger adults. Such differences need to be accounted in the design of SDS for older people. The results from JNAS corpus which has a balanced set of speakers in both the age groups and in gender, indicates that the impact on WERs with ageing is more pronounced for male speakers as compared to female speakers.

Chapter 5

Impact of changes in glottal source parameters with ageing on ASR

In chapter 4, significantly higher ASR WERs were observed for the older adults than for younger adults on the SCOTUS, MATCH and JNAS corpora. In this chapter, the differences in voice characteristics of the younger and older adult speakers are analysed and an attempt is made to delve into the possible causes for the ASR performance degradation on older voices. Several important glottal source parameters such as the fundamental frequency, jitter (measure of temporal perturbations in glottal source periods), shimmer (measure of amplitude perturbations in glottal source periods), and harmonicity for the two age groups are compared and wherever the measures differ significantly, the effect of changes in these parameters on ASR accuracies has been analysed.

5.1 Experimental setup

Among the two English corpora used in this thesis, the number of utterances available in the MATCH corpus are quite limited and not quite sufficient for detailed analysis of voice characteristics. Hence the SCOTUS corpus has been used for this set of experiments. Since the number of female speakers in this corpus is also very small, we used only the male speakers test set as described in section 3.3.1 for voice analysis. This also helps to keep the analysis free from gender induced variations. We have analysed and compared the samples of phoneme 'aa' from adult and older male speakers.

Voice analysis is typically carried out on sustained vowel pronunciations recorded in a controlled noise-free environment. However the SCOTUS corpus is spontaneous

speech with a considerable amount of background noise. Being spontaneous in nature, the corpus also does not have sustained vowel pronunciations with durations over a few seconds. Most of the samples of the vowels are typically a fraction of a second long and are part of a longer utterance. In order to pick the best available instances of the phoneme 'aa' from the speech the following procedure was used.

1. Each utterance was force aligned to the triphone transcription, in order to determine the frame boundaries and the likelihood of each triphone in the utterance.

2. All the triphone samples with the centre phoneme 'aa' were selected.

3. Out of the selected samples, the ones with negative log likelihood greater than a threshold of 1000 were rejected.

4. From the remaining, those samples having a duration less than 0.1 seconds were rejected, to get the final set of vowel 'aa' samples for analysis.

In all, 2970 samples of 'aa' from 23 adult male speakers and 2105 samples from 10 older male speakers were used for voice analysis. Several voice parameters such as the fundamental frequency, jitter, shimmer and harmonicity measures were computed for the selected samples using 'Praat' [Boersma, 2001].

Apart from these parameter computations on sustained vowels, using complete speech utterances cepstral peak prominence measures were also computed and analysed.

Each of the following sections deals with one voice parameter analysing if there is a significant difference in the parameter value between adult and older speakers. Wherever the difference is significant, we artificially modify those parameters in speech from younger adults to analyse the impact on ASR accuracies.

5.2 Fundamental frequency

The result of the analysis of fundamental frequency are tabulated in Table 5.1. We observe that the average fundamental frequency for older males is about 15 Hz (10%) lower than that of adult male voices. The differences in F_0 measures are statistically significant at $p < 0.001$ using Mann-Whitney rank sum test.

In order to understand the affect of reduction in $F0$ on ASR accuracies, we artificially reduce the F_0 by 10% and compare the WERs of the original waveforms and

F0	Younger Males		Older Males		p-value
	Mean	Std	Mean	Std	
Median F0	144.4	44.3	128.2	45.4	< 0.001
Mean F0	143.9	43.2	128.0	44.6	< 0.001

Table 5.1: Fundamental frequency analysis for the phonations of vowel 'aa' in the SCOTUS corpus.

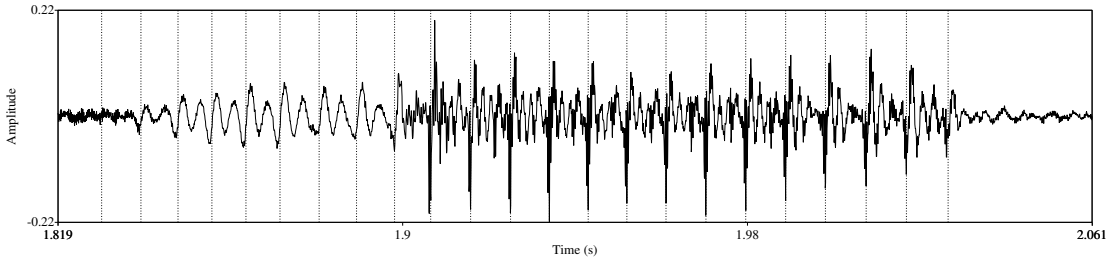
modified waveforms. The factor of 10% was used to reflect the difference in adult and older voices. For this experiment, the ASR system is the same as that described in section 4.2.1.1. We use 400 utterances from 8 adult speakers (4 Male and 4 Female) as the test set. For each waveform, the pitch tier is calculated using using Praat. The frequencies are then scaled to 0.9 of their original value. Using the new pitch tier, the waveforms are resynthesized using pitch synchronous overlap and add (PSOLA) method [Moulines and Charpentier, 1990]. Figure 5.1 shows an example of the waveforms and F_0 contours before and after $F0$ manipulation.

The word error rates before and after reduction in $F0$ are given in Table 5.2. The WER increases by 1.1% absolute to 33.2% and is statistically significant with $p < 0.001$ using the Matched pair sentence segment word error (MAPSSWE) test [Gillick and Cox, 1989]. In order to be able to attribute the increase in WER to the change in fundamental frequency and not to the resynthesis process, we repeated the resynthesis process described above without modifying the pitch tier. The WER for the resynthesized waveforms is 32.0% and the difference with respect to the original waveform is statistically insignificant with p 0.61 using MAPSSWE test.

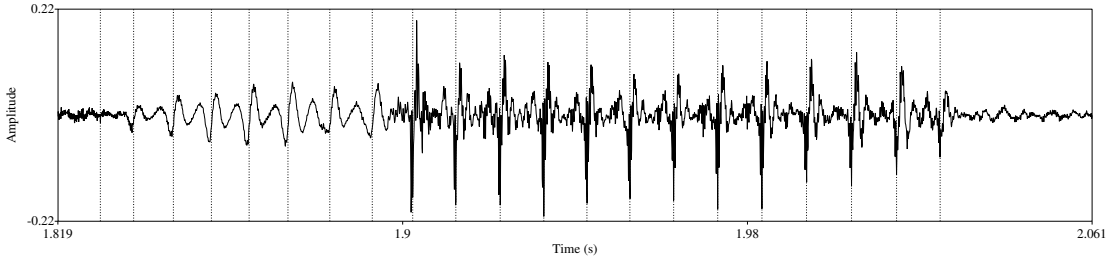
We also perform VTLN, calculating the warping factors for each speaker separately for the two sets. Using VTLN, the difference in WER is reduced to 0.7% absolute at $p < 0.01$ using MAPSSWE test.

Word Error Rate (WER) %			
	Original	Reduced pitch	p-value
Without VTLN	32.1	33.2	< 0.001
with VTLN	28.8	29.5	< 0.01

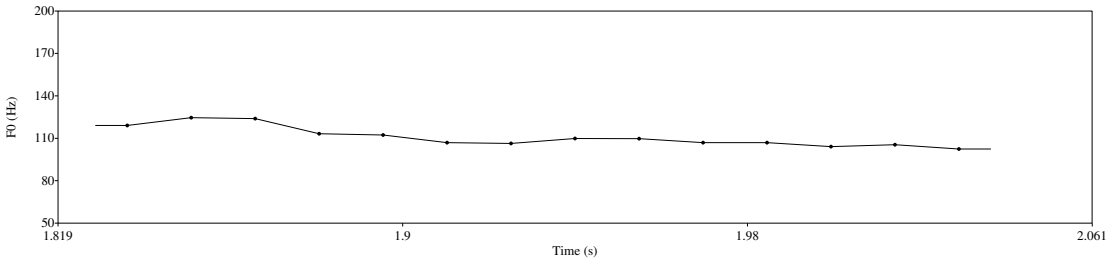
Table 5.2: WER (%) with artificial reduction in fundamental frequency of the speech from younger adults in the SCOTUS corpus.



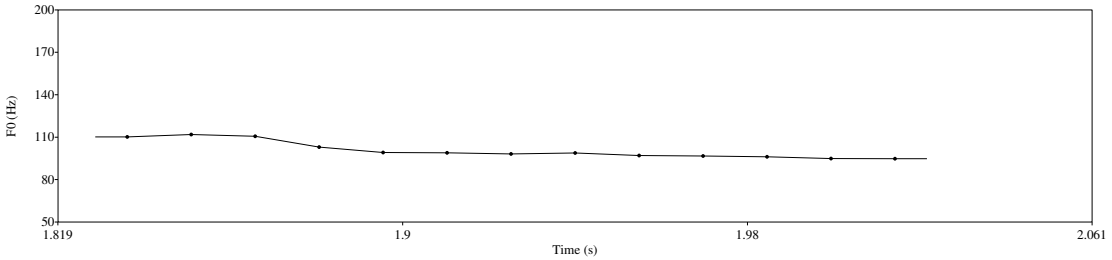
(a) Original: Waveform



(b) F_0 Modified: Waveform



(c) Original: F_0 contour



(d) Modified: F_0 contour

Figure 5.1: Illustration of artificial modification of fundamental frequency

5.3 Jitter

For our analysis, the jitter measurements 'Jitter Local (Jit Loc)' and 'Jitter Relative Average Perturbation (Jit RAP)' as described in section 2.3.3 were computed. Since the analysed samples were obtained from continuous speech, the duration of each sample is quite short. As a result each sample only has a few cycles of glottal periods. Hence the higher order measures of Jitter which average on larger number of cycles are unreliable in our experimental setup and thereby omitted in the analysis.

The variations of these jitter measurements are shown in Table 5.3. The changes are statistically significant at $p < 0.001$ using Mann-Whitney rank sum test.

Jitter	**Younger Males**		**Older Males**		**p-value**
	Mean	**Std**	**Mean**	**Std**	
Jit Loc	1.89	1.50	2.41	1.83	< 0.001
Jit RAP	0.85	0.96	1.08	1.14	< 0.001

Table 5.3: Jitter analysis for the phonations of vowel 'aa' in the SCOTUS corpus.

In order to understand the affect of increased jitter on ASR performance, we artificially introduce jitter into the 400 test waveforms from 8 speakers.

Pulse positions representing the glottal closures are extracted from the speech utterances. Each pulse position PP_{old} is then perturbed to get a new pulse position PP_{new} as follows

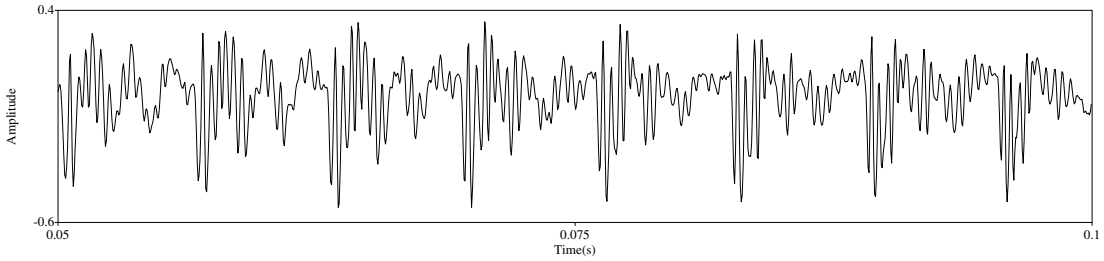
$$PP_{new} \quad PP_{old} \quad r \quad \alpha \quad T_{avg} \tag{5.1}$$

where, $-0.5 \leq r \leq 0.5$ is a uniformly distributed random variable, α is a factor controlling the maximum perturbation allowed as a fraction of the average period T_{avg}.

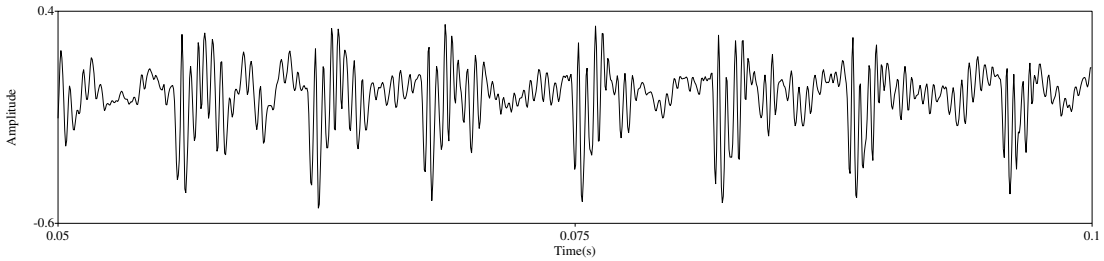
Using these new pulse positions, the waveform is resynthesized by pitch synchronous overlap and add method to get a waveform with increased jitter. Figure 5.2 shows an example of the waveforms before and after artificial increase in jitter.

Temporal perturbations with $\alpha \quad 0.05$ and $\alpha \quad 0.10$ were introduced into the waveforms. To get an idea of the jitter values before and after the modification, the same approach as explained in section 5.1 was used to sample 401 occurrences of the phoneme 'aa' from the test utterances. The jitter measures on the samples from original and modified waveforms are presented in Table 5.4.

Table 5.5 shows the ASR WERs on the original waveforms and the waveforms with increased jitter. The change in WER with increased jitter is statistically insignificant



(a) Waveform: Original



(b) Waveform: Jitter Added

Figure 5.2: Illustration of waveforms with artificial increase in jitter

Jitter	Original		α 0.05		α 0.10	
	Mean	Std	Mean	Std	Mean	Std
Jit Loc	1.63	1.41	2.31	1.52	3.08	1.69
Jit Rap	0.70	0.78	1.02	0.94	1.39	1.06

Table 5.4: Jitter values computed on phonations of the vowel ‘aa’ in the original and modified waveforms

(using MAPSSWE test) and the ASR system performance is seen to be quite robust to jitter variations.

Word Error Rate (WER) %		
Original	α 0.05	α 0.10
32.1	32.2 p 0.62	32.4 p 0.17

Table 5.5: WER (%) with artificial increase of jitter in the speech from younger adults in the SCOTUS corpus.

5.4 Shimmer

Shimmer measures 'Shimmer Local (Shim Loc)' and 'Shimmer Three point Amplitude Perturbation Quotient (Shim APQ3)' were computed using Praat. Again, due to the short duration of analysed samples, shimmer measures that are averaged over larger number of cycles have not been compared.

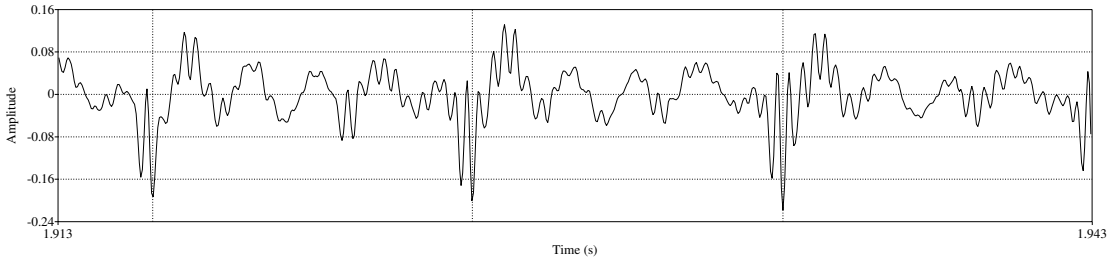
Table 5.6 shows that the shimmer measures for older males are higher compared to the adult males and the results are statistically significant (with $p < 0.001$ using Mann-Whitney rank sum test).

Shimmer	**Younger Males**		**Older Males**		**p-value**
	Mean	**Std**	**Mean**	**Std**	
Shim Loc	10.73	5.22	11.33	5.27	< 0.001
Shim APQ3	4.65	2.70	4.93	2.88	< 0.001

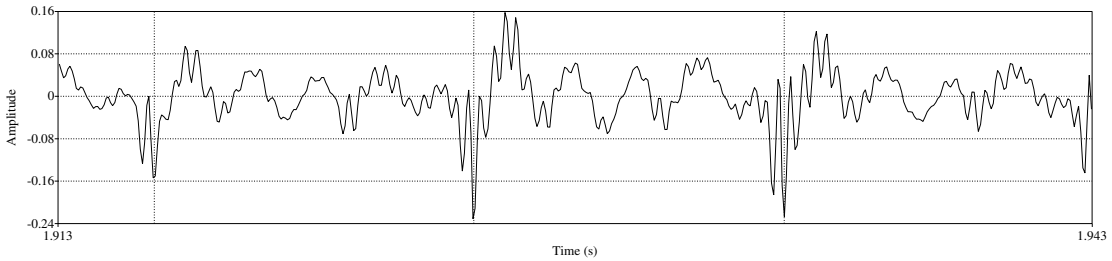
Table 5.6: Shimmer analysis for the phonations of vowel 'aa' in the SCOTUS corpus.

We artificially increase shimmer in the test waveforms to understand the affect of increased shimmer on ASR performance. Pulse positions representing glottal closures are extracted for each test waveform. From the location of the pulse positions, the voiced and unvoiced segments in speech are determined. To simulate shimmer effect, the speech samples x_{old} between two adjacent pulses in voiced segment are scaled to obtain x_{new} as follows

$$x_{new} \quad x_{old} \quad 1 \; r \quad \alpha \tag{5.2}$$



(a) Waveform: Original



(b) Waveform: Shimmer Added

Figure 5.3: Illustration of waveform with artificial increase in shimmer

where, $-0.5 \leq r \leq 0.5$ is a uniformly distributed random variable which is fixed for all the speech samples between two adjacent pulses, and α is a factor controlling the maximum perturbation allowed.

An example of the waveform and spectrograms before and after artificial introduction of shimmer is seen in Figure 5.3. Similar to the Jitter measurements in Table 5.4, Shimmer values measured over the 401 segments of phoneme 'aa' in the test utterances before and after artificial increase of shimmer are presented in Table 5.7.

Shimmer	Original		α 0.05		α 0.10	
	Mean	Std	Mean	Std	Mean	Std
Shim Loc	9.71	5.43	10.33	5.41	11.12	5.44
Shim APQ3	3.94	2.72	4.25	2.74	4.76	2.73

Table 5.7: Shimmer values computed on phonations of the vowel 'aa' in the original and modified waveforms

Table 5.8 shows the results with maximum perturbation in amplitude between adjacent periods of 5% and 10%. The effect of shimmer on ASR WERs is seen to be insignificant.

Word Error Rate (WER) %		
Original	α 0.05	α 0.10
32.1	32.1 p 0.65	32.1 p 0.13

Table 5.8: WER (%) with artificial increase of shimmer in the speech from younger adults in the SCOTUS corpus.

5.5 Harmonicity

For the measurement of parameters indicating breathiness, autocorrelation (Autocorr) and Noise to Harmonic Ratio (NHR) were computed from the chunked 'aa' segments. CPP and CPPS were also measured using the whole speech utterances instead of chunked phoneme utterances. The results are tabulated in Table 5.9.

Harmonicity	**Younger Males**		**Older Males**		**p-value**
	Mean	**Std**	**Mean**	**Std**	
Autocorr	0.85	0.08	0.85	0.09	0.61
NHR	0.21	0.15	0.21	0.16	0.79
HNR (dB)	9.03	3.15	9.10	3.16	0.49
CPP	10.81	0.83	10.69	0.82	$<$0.001
CPPS	2.71	0.43	2.69	0.4	$<$0.05

Table 5.9: Harmonicity analysis for the phonations of vowel 'aa' in the SCOTUS corpus.

It is observed that the differences in the harmonicity measures of younger adult and older adult males are statistically insignificant (by Mann Whitney rank sum test). Though the changes in CPP and CPPS measures are found to be statistically significant, the actual difference in the values is very small. CPPS which has been reported by Hillenbrand and Houde [1996] to be better correlated with perceived breathiness in voice than CPP, differs only by a value 0.02 for the two age groups. This coupled with the comparative results of NHR suggests that the difference in the breathiness characteristics of younger and older male test sets used in our experiments do not differ much.

5.6 Summary

Many of the values of the voice analysis measures reported in this article are somewhat higher than the published values in diagnostic medical research. This is due to the fact that we have not used sustained vowel pronunciations in clean recording conditions, but extracted sustained phonemes from spontaneous speech. Due to chunking, there is also a co-articulation effect at the beginning and the end of each analysed phoneme sample. However the same procedure has been applied to both adult and older voices in similar recording environments to analyse the differences between the two groups. Indeed our analysis is relevant in this context as it is made on natural speech which is the typical input to ASR systems.

Jitter and Shimmer measures have been extensively studied and have been used by researchers in age recognition from voice. From our experimental results too, we observe a clear increase in jitter and shimmer values for older voices. These measures can work well for detection of ageing voices. However, the variations in these measures do not have a significant impact on ASR accuracies. Front end feature extraction techniques in ASR such as perceptual linear prediction used in our experiments are quite robust and suppress the variations in the glottal source characteristics.

Changes in the fundamental frequency appear to increase the errors marginally, which can be overcome to some extent using vocal tract length normalisation.

The speech from older adults used in our experiments do not show a significant increase in parameters related to breathiness. It is however an important parameter that needs to be further investigated.

Chapter 6

Articulatory changes in older voices

As observed in the previous chapter, although there are significant differences in the glottal source characteristics of younger and older adults, these changes do not contribute significantly towards the reduction in ASR accuracies. In this chapter, some of the aspects of the changes in articulation patterns with ageing are studied. In particular, it is of interest to see which phonemes are most affected in terms of recognition accuracies. Phoneme accuracies on SCOTUS corpus and JNAS corpus are analysed to see if any patterns emerge across corpora and across speakers.

Another widely studied articulatory parameter in vocal ageing research is the 'rate of speech'. Speaking rate has been reported to be slower in older adults as compared to younger adults. However, the impact of slower speech rate on ASR accuracies is not well understood. This issue is also addressed in this chapter with experimental analysis.

6.1 Phoneme recognition accuracies

As discussed in section 2.2.3, several changes have been reported in the physiology of the articulators with ageing. These include restricted jaw movement, loss of tongue strength and the rate of movement of these articulators. This results in changes in articulatory patterns during old age.

An interesting question that needs to be answered is whether these changes impact all the phonemes in terms of ASR accuracies. Typically the hypotheses generated by the ASR system are constrained by the allowed pronunciations imposed by the lexicon and the sequence of words allowed by the langauge model. Hence simply expanding the decoded word hypothesis to phoneme level hypothesis using the lexicon will not

lead to proper insights.

In order to analyse the results at phoneme level, we remove all such confounding affects by using a phoneme loop decoder. A phoneme loop decoder is a finite state machine which allows any phoneme to follow any phoneme as shown in Figure. 6.1. Under such an unconstrained setting, the overall results in terms of percentage of correct recognitions and overall accuracies are usually much lower than the results obtained using language models.

In the following experiments, we consider phoneme correct recognition percentage as an evaluation metric. It is a ratio of number of correct recognitions of a phoneme in the decoded hypothesis to the total number of occurrences of the phoneme in the reference transcript. The results are computed after dynamic programming based string alignment of the reference and decoded hypotheses. Such an alignment procedure allows the comparison of corresponding reference and recognition labels for a given speech segment and hence the computation of accuracies and phoneme confusion matrices.

It is also important to note that the usage of phonemes in a language does not have a uniform distribution. Hence while analysing the overall impact on ASR accuracies, the probability of the phoneme in the language is used to weigh the difference in phoneme error for that phoneme between the two age groups.

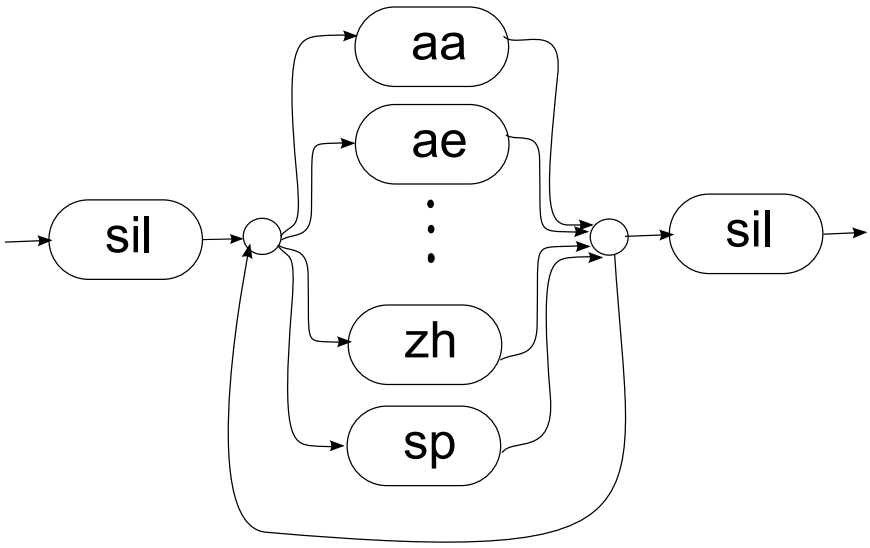


Figure 6.1: Phoneme loop decoder

6.1.1 Results on the SCOTUS corpus

For the experiments with SCOTUS corpus, the training set and the test set are same as those used in section 4.2.1.1. Since the number of older female adults are quite limited, we analyse the phoneme errors of only the male speakers in this corpus. Monophone HMMs were trained without any state tying. Each phoneme was modeled as a three state HMM with 18 Gaussian components per state.

Figure. 6.2 shows the comparative results of correct recognition of each phoneme for the two age groups. The results are categorised based on the phonetic classes and the actual numbers can be found in the appendix in Table A.10. Only a few phonemes are seen to have drastic reduction in correct recognitions. Among the monophthongs, the lower vowels (*aa, ao, ae*) seem to be the most affected with over 10% difference between the two age groups. All the diphthongs show comparable results except *aw* for which the recognition drops by over 10%. In consonants, the fricative *hh* has a substantial drop in performance. The r-coloured vowel *er* is the other phoneme that sees a drop of over 10%. The nasals (*m, n, ng*) have about 3-5% decrease in recognition rates.

To understand which phonemes have the most impact on overall increase in phoneme error rates, the differences in phoneme correct recognition between the age groups are scaled by the probability of the phoneme occurrence in the language. The phoneme statistics are computed over all the utterances in the SCOTUS corpus and shown in Table A.10.

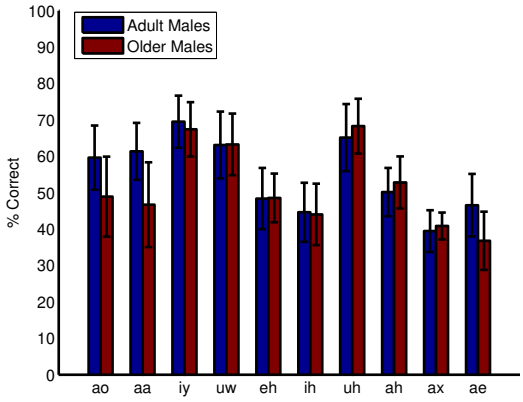
The phonemes with most dominant affect on ASR accuracies in descending order on SCOTUS corpus are as follows:

ae, aa, er, t, n, ao

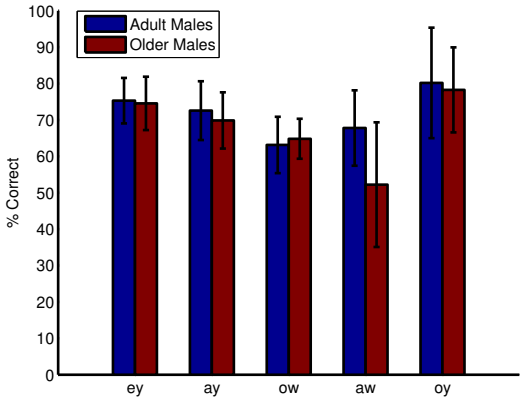
6.1.2 Longitudinal results on the SCOTUS corpus

In section 4.2.1.2, it was seen that ASR accuracies deteriorate longitudinally for older speakers in the SCOTUS corpus. In this section we analyse the phoneme recognition rates longitudinally for those speakers. The motivation behind this experiment is to see if patterns emerge in phoneme errors across speakers.

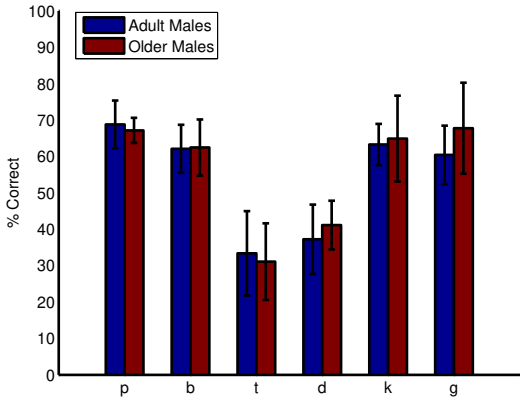
Using a phoneme loop decoder, test utterances from 5 adult male speakers from the SCOTUS corpus were decoded. The same monophone acoustic models as described in section 6.1.1 are used. For each of the speakers the test set comprises of about 200 utterances each recorded about 8 years apart. Table 6.1 shows the phonemes that have



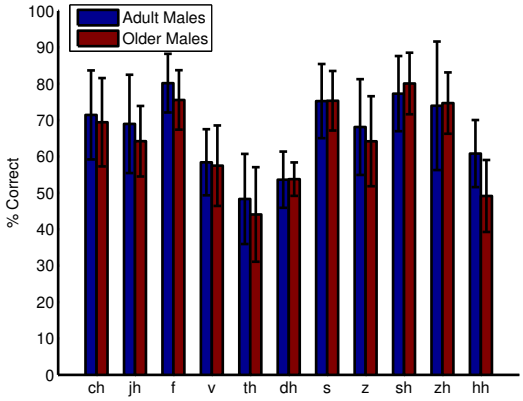
(a) Monophthongs



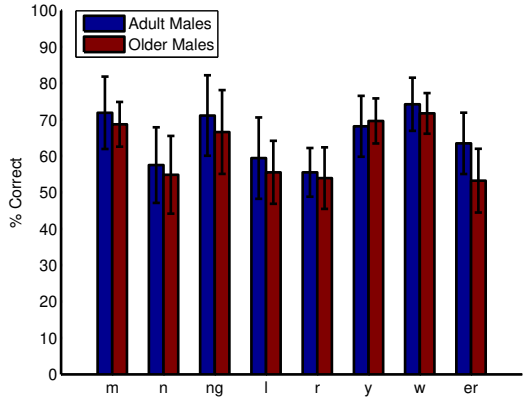
(b) Diphthongs



(c) Consonants



(d) Affricates (ch, jh) and Fricatives (f, v, th, dh, s, z, sh, zh, hh)



(e) Nasals (m, n, ng), Liquids (l, r), Semi vowels (y, w) and R Coloured vowel (er)

Figure 6.2: Phoneme correct recognition (%) on the SCOTUS corpus

more than 10% decrease in phoneme recognitions with ageing. The phonemes in the table for each speaker are sorted in decreasing order of the difference in recognition rate longitudinally.

Speaker Id	Start age	End age	Phonemes with largest decrease in recognition rate
02	79	87	*g, jh, ch, aa, eh, l, r, hh*
03	63	71	*aw, g, zh*
04	62	70	*oy, s, ch, b, th*
05	59	67	*oy, uh, n, jh*
08	60	68	*zh, oy, aa*

Table 6.1: Phonemes with largest drop in recognition rates in longitudinal study on the SCOTUS corpus

It is seen that there is a large variability in terms of most affected phonemes across speakers.

6.1.3 Results on the JNAS corpus

In the baseline results on the JNAS corpus in section 4.2.3, it was seen that the difference in WERs is highest between younger adult males and the older adult males in the age range of 70-79 years. We use these two sets to compare the differences in phoneme recognition rates. Similar to the experimental setup for the SCOTUS corpus, monophone HMMs with 16 Gaussian components per state were trained, and a phoneme loop decoder experiment was setup for JNAS corpus. The phoneme recognition rates of the two test sets was compared.

Analysis of the monophone transcripts of the JNAS transcripts suggest that some phonemes occur quite a lot while the occurrence of certain phonemes is negligibly low as seen in table A.11. Hence from an ASR point of view, we look at the phonemes that occur the most. Figure 6.3 shows the comparative recognition rates of younger male and older male speakers in descending order of occurrence.

On scaling the differences in phoneme recognition rates with the probability of occurrence of the phonemes, the phonemes that appear to have a major impact on the overall decrease in accuracies for older adults are as follows:

i, a, r, e, s, m, u

Comparing the results on the SCOTUS and JNAS corpora, some of the lower vowels seem to be commonly affected by ageing.

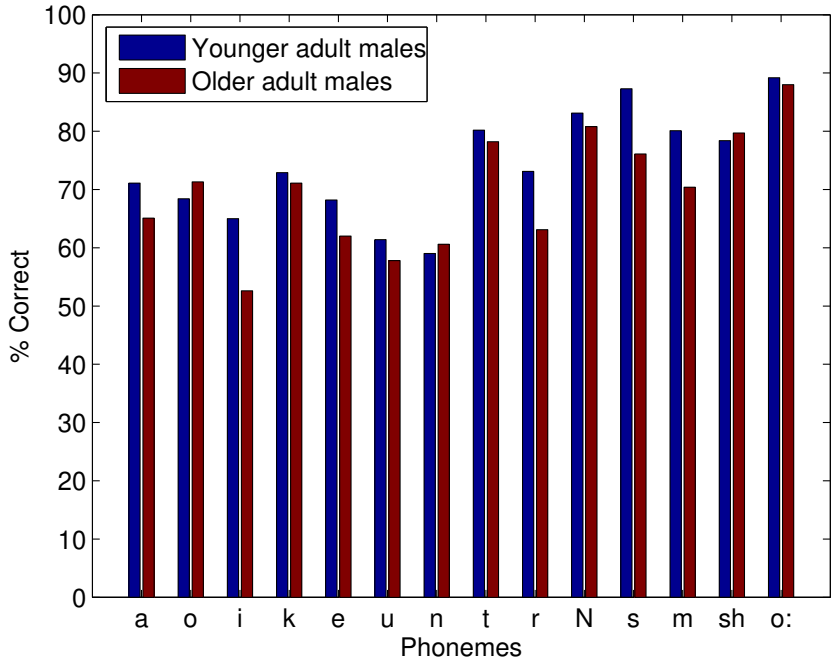


Figure 6.3: Correct recognition (%) of most used japanese phonemes for the younger and older adults. Refer Table. A.11

6.2 Vowel centralisation

If the movement of the articulators are restricted in terms of force and range, it results in an undershoot of vowel articulation. This undershoot can lead to changes in the formant patterns, viz., formants with higher frequency tending towards lower frequency and formants with lower frequencies tending towards higher frequencies. This effect is called vowel centralisation. It has been reported by Liss et al. [1989] that vowel centralisation is quite pronounced in very old speakers with all vowel realisations sounding quite close to each other.

Vowel centralisation is typically measured using the vowel space area. First and second formant frequencies (F1 and F2) are calculated for each vowel and the vowels are plotted in the 2 dimensional F1-F2 space. The vowel space area is the area enclosed by the corner vowels *i, u, a* and *ae*.

For the vowel space analysis, speech samples from the SCOTUS corpus were used. The analysis was again carried out on male speakers due to the limitation in the number of female speakers. The utterances used for analysis were the same as those used in section 5.1 and the voice samples for vowel space analysis were chosen in similar manner.

1. Each utterance was force aligned to triphone transcription, in order to determine the frame boundaries and the likelihood of each triphone in the utterance.

2. All the triphone samples with central vowel phoneme were selected.

3. Out of the selected samples, the ones with log likelihood/frame less than a threshold of -80 and with a length less than four frames (40ms) were rejected.

Using 'Praat', the values of first (F1) and second formant (F2) frequency for each vowel instance were computed at the midpoint of that instance's duration. For each vowel for a speaker, lower quartile (LQ), upper quartile (UQ) and interquartile range (IQR = UQ-LQ) were computed. Outliers outside the range [(LQ-1.5IQR), (UQ+1.5IQR)] were rejected. Mean values of each vowel for a speaker were computed from the remaining samples.

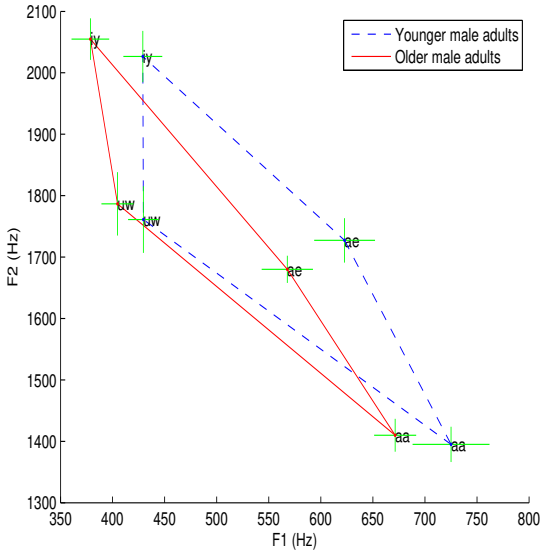


Figure 6.4: Mean vowel space areas for younger and older male adults in the SCOTUS corpus. Corner vowels and their standard deviations are also shown in the figure.

The vowel space bounded by the phonemes *aa, uw, iy* and *ae* for both the age groups is shown in figure 6.4. The corner points of each quadrilateral is the average across all the speakers in that age group. The area of the vowel quadrilateral for each speaker is computed by summing the areas of the triangles formed by the points *iy, uw, ae* and *aa, uw, ae* for that speaker. The area of the triangles is in turn calculated using Heron's formula

$$Area \quad \sqrt{s\ s-a\ \ s-b\ \ s-c} \tag{6.1}$$

where, a, b, c are the sides of the triangle and $s \quad a \ \ b \ \ c \quad 2$

It is seen from Table 6.2 that the area occupied by the vowel quadrilateral of older speakers is less than that of younger speakers, indicating vowel centralization. The vowel space areas of the speakers in the two age groups are significantly different at $p < 0.01$ using student T test.

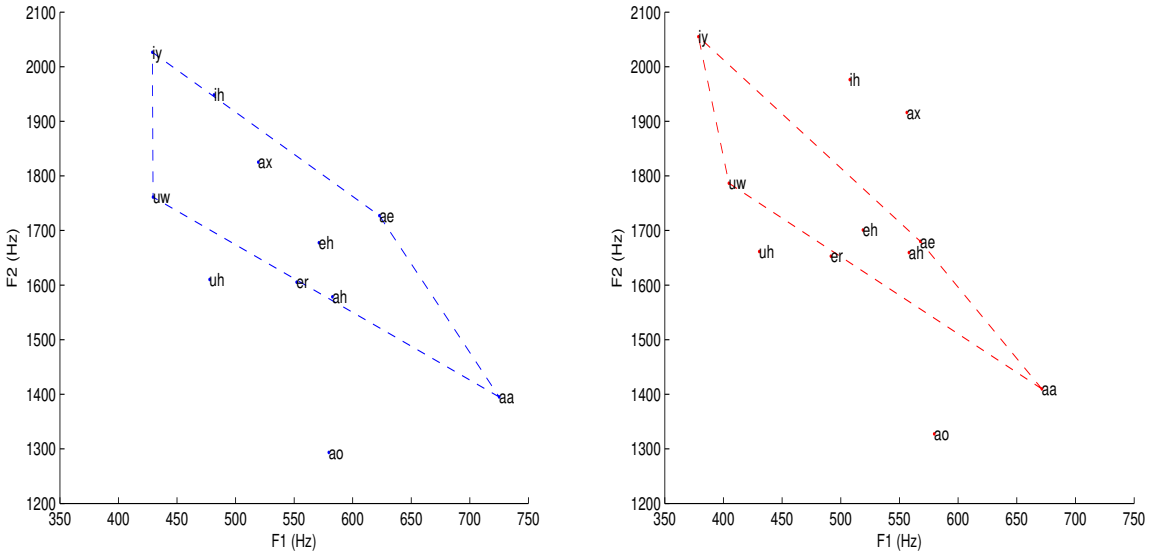
Vowel space area Hz^2)			
Younger adult males	**Older adult males**	**Difference**	**p-value**
5.46 10^4 (std: 1.33 10^4)	3.99 10^4 (std: 0.97 10^4)	1.47 10^4	< 0.01

Table 6.2: Vowel Space Area comparison between *younger adult* and *older adult* males in SCOTUS corpus

While all the corner vowels appear to shift in the F1-F2 space with ageing, the phoneme recognition results reported in Figure 6.2 show that vowels 'ae' and 'aa' have a large decrease in performance with ageing while phonemes 'iy' and 'uw' do not show much difference in accuracies.

In order to understand this better, the centroids of all the monophthongs (averaged over all the speakers in each age group) are shown in Figure 6.5. For the older speakers, the vowels appear to move closer to each other into clusters in the F1-F2 space especially in the central region of the vowel space. This might lead to a reduction in the discrimination capacity between phonemes (atleast in the F1-F2 space) and also explain to some extent the large decrease in recognition accuracies for some of the central vowels. It is also interesting to see that with ageing, there is a tendency of F1 decreasing and F2 increasing for most of the phonemes.

It is important to clarify at this point that in the calculation of F1 and F2 for the vowels, segments from continuous speech were used and thus may have inherent confounding co-articulation effects with adjacent phonemes. The segments chosen are also typically very short in duration and more so in case of short vowels such as 'ih', 'ax', 'eh' and 'er' which further impacts the accurate computation of the true formant frequencies. However, extreme care has been taken to remove the outliers and to choose the best samples available for analysis with the same procedures applied for both the age groups. In Figure 6.5.(b), for instance, the short vowels 'ih' and 'ax' seem to shift significantly for older age group. It is not clear if this shift is indeed due to the



(a) Vowel space: younger adults

(b) Vowel space: older adults

Figure 6.5: Centroid positions of common vowels in younger and older Adults

influence of ageing or due to the limitation of the procedure of choosing samples from continuous speech. It needs further investigation with experiments on other corpora to understand clearly this effect.

6.3 Speaking rate

Speaking rate has been reported to be lower in older adults than younger adults. Though the underlying physiological change for the decrease in speech rate is not clear completely, it is believed to be a result of restricted free movement in articulators and a reduction in motor control capabilities. It has also been suggested that older adults tend to deliberately reduce speech rate so as to be more intelligible under restricted motor control abilities.

Slower speaking rate is a combined effect of longer pronunciation of words, increased number of pauses and pause duration. The impact of speaking rate differences on ASR accuracies has received little attention in ASR research. Fosler-Lussier and Morgan [1999] indicate that the WER increases marginally with increased speaking rate.

In this section, the differences in speaking rates on SCOTUS and JNAS corpus are analysed. Two different approaches are employed to compute and compare the

speaking rates between the two age groups. The impact of speaking rate differences on ASR accuracies are then investigated on the JNAS corpus.

6.3.1 Speaking rate comparison on SCOTUS corpus

For our analysis of speaking rate, we compute the average number of frames (amount of time) per phoneme. Analysis is done on the younger and older adult male test sets described in section 5.1. The utterances being analysed were first force aligned to a phoneme transcription. All the silences and short pauses were then deleted. The purpose of deleting the pauses was to analyse if there is a difference in the speaking rate in the speech part of the utterances. The average duration (d_p) for each phoneme for each age group was then computed using the information from forced alignment results.

For overall speaking rate, a weighted average of all the phoneme durations is computed as follows:

$$d \quad \sum_{p \ P} w_p \quad d_p \tag{6.2}$$

where, w_p is the probability of occurrence of the phoneme p.

We find from our results in Table 6.3 that there is a statistically significant decrease in the speaking rates in older voices. From table A.12, it is also seen that there is a consistent decrease in speaking rate with ageing for all the phonemes.

Average duration (msec) per phoneme		
Younger Adult males	**Older adult males**	**p**
81.0	90.8	< 0.001

Table 6.3: Speaking rate differences between younger and older adults on the SCOTUS corpus

6.3.2 Speaking rate comparison on JNAS corpus

For the analysis of speaking rate on JNAS corpus, we use two different approaches:

1. using the state occupancy probabilities of the phoneme HMMs.

2. using forced alignment method as described in the previous subsection.

In the JNAS corpus, we have a reasonably large and equivalent amount of training utterances for each gender-age category viz., Male-Young, Male-Old, Female-Young, Female-Old. Using these sets, monophone acoustic models were trained for each category. The model set for each category comprises of 3 state HMMs for each phoneme with 16 Gaussian components per state.

Given an HMM with transition probabilities in state s given by $a_{ij}^{\ s}$, the expected duration d_p of occupying all the states S of the HMM for phoneme p is given by [Rabiner, 1989]:

$$d_p \quad \sum_{s \ S} \frac{1}{1 - a_{ii}^{\ s}} \tag{6.3}$$

Equation 6.3 gives the expected number of frames emitted by a HMM and thus can be used as a measure of frames/phoneme. The average number of frames occupied per phoneme in a model set P can again be similarly computed as a weighted average of durations over all the phonemes.

$$d \quad \sum_{p \ P} w_p \quad d_p \tag{6.4}$$

The weights w_p are the expected probability of a phoneme in the language. From the phoneme counts over the whole JNAS corpus, the weights were approximated and tabulated in Table A.13.

Table 6.4 shows the expected duration per phoneme for each of the models trained. It is seen that the speaking rate is slower for older adults both male and female as compared to their younger counterparts.

Speaking Rate (msec per phoneme)				
Method	**Males**		**Females**	
	Younger adults	**Older adults**	**Younger adults**	**Older adults**
Model Based	71.7	94.9	78.9	98.8
Forced alignment	72.0	91.2	78.1	93.8

Table 6.4: Speaking Rate differences between younger and older adults in the JNAS corpus with a) model based method where the transition parameters of the hidden Markov models are used to estimate the expected occupancy of each phoneme and b) using forced alignment method to compute average number of frames associated with each phoneme.

As seen in Table 6.4, similar results are obtained even with the forced alignment method as described in Section 6.3.1.

6.3.3 Impact of speaking rate changes on ASR accuracies

From the speaking rate analysis on JNAS corpus, it is observed that the acoustic models for younger and older adults differ by a large margin in transition parameters across all the phonemes. The probabilities of occupying the same state is higher in models trained on older speakers as compared to those trained on younger speakers.

Let the monophone acoustic models trained on each age-gender category be represented as follows:

Younger Male Adults : $\Theta_{YM} \quad \{\mu_{YM}, \sigma_{YM}, W_{YM}, T_{YM}\}$

Older Male Adults : $\Theta_{OM} \quad \{\mu_{OM}, \sigma_{OM}, W_{OM}, T_{OM}\}$

Younger Female Adults : $\Theta_{YF} \quad \{\mu_{YF}, \sigma_{YF}, W_{YF}, T_{YF}\}$

Older Female Adults : $\Theta_{OF} \quad \{\mu_{OF}, \sigma_{OF}, W_{OF}, T_{OF}\}$

To understand the impact of changes of duration, we replace the transition parameters of the models corresponding to younger male adults by the transition probabilities of the Older male adults. The transition parameters for the other models are also replaced similarly to get a set of modified models.

$\widehat{\Theta}_{YM} \quad \{\mu_{YM}, \sigma_{YM}, W_{YM}, T_{OM}\}$

$\widehat{\Theta}_{OM} \quad \{\mu_{OM}, \sigma_{OM}, W_{OM}, T_{YM}\}$

$\widehat{\Theta}_{YF} \quad \{\mu_{YF}, \sigma_{YF}, W_{YF}, T_{OF}\}$

$\widehat{\Theta}_{OF} \quad \{\mu_{OF}, \sigma_{OF}, W_{OF}, T_{YF}\}$

Using a phoneme loop decoder, the test sets (as described in section 4.2.3) for each age-gender category is decoded using the original and modified acoustic models for that category.

The results for older speakers are shown in Table 6.5. These results capture in effect, the outcome of slower speech test set decoded on models trained on slower speech and models trained on faster speech, all other parameters of speech being the same. It is observed that while correct recognitions are almost the same, the accuracies suffer a large decrease for both male and female speakers with modified models. It can thus be concluded that insertions errors increase for slower speech decoded with models trained on relatively faster speech.

Phoneme accuracies using phoneme loop decoder for older speakers				
Acoustic models	**Older males**		**Older females**	
	% Correct	**% Accuracy**	**% Correct**	**% Accuracy**
Original (Θ)	72.3	54.2	78.9	62.4
Modified ($\widehat{\Theta}$)	72.5	50.1	78.9	59.9

Table 6.5: Phoneme accuracies using phoneme loop decoder for older speakers

Table 6.6 captures similar results for younger speakers. It is interesting to note that insertion errors are less even for faster speech decoded on models suited for slower speech. While the correct recognition rates are marginally lower with modified models, the overall accuracies seem to be better. These results are however not constrained by language model weighting and do not completely explain the outcome in a complete ASR setup.

Phoneme accuracies using phoneme loop decoder for younger speakers				
Acoustic models	**Younger males**		**Younger females**	
	% Correct	**% Accuracy**	**% Correct**	**% Accuracy**
Original (Θ)	73.3	58.7	73.9	61.0
Modified ($\widehat{\Theta}$)	72.9	60.4	73.9	62.7

Table 6.6: Phoneme accuracies using phoneme loop decoder for younger speakers

We repeat the above experiments using a full ASR decoder including the language models and lexicon. The experimental setup is similar to that explained in section 4.2.3, except that in the current set of experiments instead of triphone acoustic models, we use monophone models trained separately for each age-gender category.

The accuracies for the older speakers are tabulated in Table. 6.7 and the split of the errors in terms of substitutions, deletions and insertions is shown in Table 6.8.

We note that while (substitution + deletion) errors remain almost the same, insertion errors for slower speech with models tuned for faster speech are 1.2% absolute higher for male speakers and 0.3% absolute higher for female speakers. It is interesting that although the reduction in speaking rate is similar for both the older males and females, there is a considerably higher insertion error rate for older males with models trained on relatively faster speech.

The results for the younger speakers are shown in Tables 6.9 and 6.10.

Word accuracies for older speakers				
Acoustic models	**Older males**		**Older females**	
	% Correct	**% Accuracy**	**% Correct**	**% Accuracy**
Original (Θ)	75.8	69.6	84.3	80.0
Modified ($\widehat{\Theta}$)	75.6	68.1	84.2	79.6

Table 6.7: Word correct recognition and accuracies for older speakers in the JNAS corpus with original and transition parameter modified models

Word Error details (%) for older speakers						
Acoustic models	**Older males**			**Older females**		
	Subs	Dels	Ins	Subs	Dels	Ins
Original (Θ)	21.0	3.2	6.3	13.8	1.9	4.3
Modified ($\widehat{\Theta}$)	21.6	2.8	7.5	14.0	1.8	4.6

Table 6.8: Substitution, deletion and insertion errors for older speakers in the JNAS corpus with original and transition parameter modified models

Word accuracies for younger speakers				
Acoustic models	**Younger males**		**Younger females**	
	% Correct	% Accuracy	% Correct	% Accuracy
Original (Θ)	77.8	73.1	80.7	77.4
Modified ($\widehat{\Theta}$)	77.2	72.9	80.5	77.2

Table 6.9: Word correct recognition and accuracies for younger speakers

Word Error details (%) for younger speakers						
Acoustic models	**Younger males**			**Younger females**		
	Subs	Dels	Ins	Subs	Dels	Ins
Original (Θ)	19.0	3.3	4.7	16.6	2.7	3.4
Modified ($\widehat{\Theta}$)	19.1	3.6	4.3	16.6	2.9	3.3

Table 6.10: Substitution, deletion and insertion errors for younger speakers in the JNAS corpus with original and transition parameter modified models

In the test on younger speakers, once again the insertion errors reduce with models tuned for slower speech but only marginally by 0.4% absolute for males and 0.1% absolute for females. However, the deletion errors increase by 0.3% and 0.2% for males and females with modified acoustic models. Overall, there is no evidence of a large difference in errors for faster speech decoded on models trained on faster speech and models trained on slower speech.

6.4 Summary

In this chapter we have looked at the increase in ASR errors with ageing from the point of view of articulatory changes.

Comparative analysis of the phoneme errors on SCOTUS and JNAS corpus between the two age groups suggests that certain lower vowels have higher increase in errors with ageing. However, it is difficult to find strong patterns and generalise age related disfluencies across speakers as seen from the longitudinal results on the SCOTUS corpus.

The study of vowel space area changes shows vowel centralisation with ageing where the vowels move closer to each other in the first and second formant space.

The experiments to analyse the impact of slower speaking rate as observed in older speakers on ASR accuracies suggest an increase in insertion errors and the impact is seen to be more dominant for older male speakers.

Chapter 7

Acoustic models for older voices

The main concern of this thesis is to improve the acoustic models for older voices. Hence we are interested in addressing the question whether speech from older speakers is different from that of younger speakers in the acoustic space. We seek to answer the question through a speaker age-group classification task. Results of the speaker classification using two different approaches are presented. Motivated by the classification accuracies, we then look at ASR accuracies using supervised hierarchical models based on gender and age. Hierarchical models are then built in an unsupervised manner and the ASR accuracies analysed. Finally we look at a simple strategy based on speaking rate differences to refine the acoustic models for older speakers.

7.1 Speaker classification and clustering

In order to understand how close or separable the acoustic features of the older voices are from those of the younger voices, experiments on speaker clustering were performed. Since the objective is to understand the effect of ageing on the features used in the ASR system, the experiments are based on MFCC and PLP co-efficients. Prosodic features that can give a good discrimination for age based clustering as seen in chapter 5, have not been used in these experiments.

The MATCH corpus was used in these experiments since it has a roughly balanced set of 24 younger and 26 older speakers. In the first set of experiments, Support Vector Machine (SVM) classifiers were trained for younger and older voices and the classification accuracies measured. In the second set of experiments, using MLLR transforms as feature vectors, the speakers were clustered into four groups using repeated bisections.

7.1.1 Age group classification using SVMs

The goal of this exercise is to see how well the speakers can be classified into their respective age groups automatically using a simple Support Vector Machine (SVM) based classifier.

MFCCs were computed from the utterances of all the younger and older speakers. A window size of 25 ms and a frame shift of 10 ms was used in the feature extraction to get 14 dimensional vectors (including the energy).

SVMs with Radial Basis Function (RBF) kernels were used for classification. In SVM approach of classification, the data is mapped into a higher dimensional space and a linear separating hyperplane with maximal margin is found in the higher dimensional space. The RBF kernel does a non-linear mapping of the data into higher dimensional space and takes the form

$$K\ x_i, x_j \quad exp\ -\gamma\ x\ \ _i - x_j\ ^2 \tag{7.1}$$

For the classifier to function properly, two parameters viz γ - the RBF kernel parameter and C the penalty parameter for the error term, need to be fixed apriori. A cross validation with a grid search on the parameters was performed and the values of the parameters that gave the best performance were fixed at: $C \quad 256$ and $\gamma \quad 2$.

Since the number of speakers in the corpus is limited, a 'Leave one out' approach was used. The SVM classifiers were trained using LIBSVM [Chang and Lin, 2001]. The following steps were involved in the process:

1. For each test speaker, training set from the rest of the speakers' data was created.

2. All the training vectors were normalised (to zero mean and unit variance) and the same normalisation applied to the test set.

3. Using a stratified selection approach, a subset of 10000 training samples and 2000 test samples were selected. This was done since the number of training samples was very high making it computationally intensive to train the models.

4. Binary classifiers were trained for each test speaker, the two classes being 'younger' and 'older'

5. The class for each sample in the test set was predicted using the classifiers and the accuracies calculated.

Each speaker was assigned to a class based on majority vote. The overall accuracy of the classification was 70%. The precision and recall for each class are shown below in Table 7.1.

Older speakers		**Younger speakers**	
Precision	**Recall**	**Precision**	**Recall**
73.9%	65.4%	66.7%	75%

Table 7.1: Precision And recall for each class in age group classification task on MATCH corpus using support vector machines

From the results, it is seen that variations in speech due to ageing reflect in the feature vectors used in ASR and it is possible to estimate the age group of a speaker with reasonable accuracy. This 2 class problem is a relatively simpler task than age recognition.

These accuracies are consistent with the accuracies obtained in speaker age and gender recognition literature. Metze et al. [2007] report an average accuracy of around 45% with state-of-the-art systems when identifying a speaker age category on a relatively more challenging task of 7 classes comprising of children, young males, young females, adult males, adult females, elderly males and elderly females respectively. Human recognition on the same task had a precision of 54.7% and recall of 69.3%. In a perceptual study on age recognition by voice, Schötz [2001] reports that there is a better than chance probability of human listeners judging the speaker age within 10% of chronological age.

7.1.2 Speaker clustering based on MLLR transforms

MLLR transforms used in speaker adaptation map the means of speaker independent HMMs to fit the target speaker more closely. These transforms can be used as speaker identity and have been used in speaker recognition tasks [Stolcke et al., 2005].

We propose a new metric to calculate the distance between two speakers using MLLR transforms as feature vectors. The metric is explained in depth in chapter 8. In brief, the distance d_{TS} between two speakers whose MLLR transforms are represented by A_T and A_S is given by:

$$d_{TS} \quad \sum_{k\ 1}^{K} \frac{A_T - A_S\, c_k}{c_k} \tag{7.2}$$

where c_k is the k^{th} mean of the K clusters computed from the means of all Gaussians in the speaker independent model.

Using the AMI speaker independent models (described in section 3.3.2), MLLR transforms were computed for all the speakers in the MATCH corpus. To compute the distance between each of these speakers using equation 7.2, 1000 points (c_k) were computed from the means of all the Gaussians in SI model. The similarity between the speakers was then computed as $similarity \quad 1000 - d$.

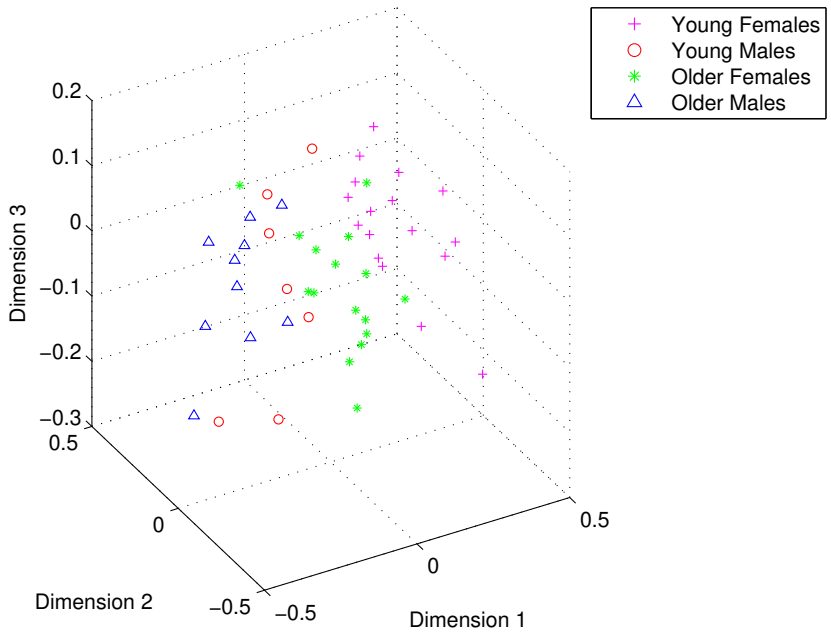


Figure 7.1: MATCH speakers in 3D space using multi dimensional scaling on the distance matrix

Figure 7.1 shows a 3 dimensional plot of all the speakers in the MATCH corpus. The plot was generated using multi dimensional scaling on the distance matrix between all pairs of speakers.

For the classification task, the speakers were clustered into four groups using CLUTO [Karypis, 2003] by the repeated bisections method. In this method, the speakers are first clustered into 2 groups, which are again bisected to obtain the desired number of clusters. The bisection is based on maximising the objective function as shown in equation 7.3

$$maximise \sum_{i=1}^{N} \sqrt{\sum_{v,u \in S_i} sim\ u, v} \tag{7.3}$$

where, N is the total number of clusters, S_i is the set of speakers assigned to cluster i, u and v represent the two speakers in that set, and $sim\ u, v$ is the similarity between the two speakers.

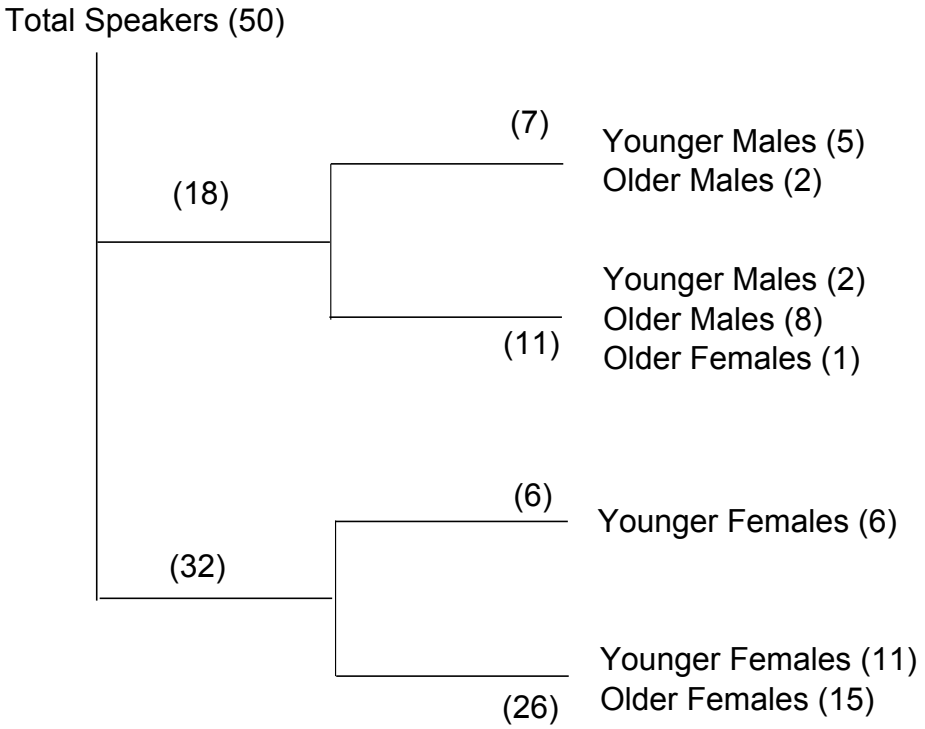


Figure 7.2: Clustering of speakers in the MATCH corpus

The speaker distribution in the clusters is shown in Figure 7.2. As expected at the first level of clustering, the male and female speakers are separated out. At the next level of clustering, there appears to be some separation between the younger and older male speakers. While for the females, there is a large overlap of younger and older females in a cluster. This also corroborates with the fact that age related changes in the voice for women are less pronounced than those observed in men.

The overall accuracy of the system is 68%. One advantage of this method over the previous SVM based method is that there is no need to tune any parameters.

7.2 Supervised hierarchical models

Speaker independent models have to generalise in order to cater to the large variety in speaker space. When the models are built for targeted sets of speakers, the recognition

accuracies are higher. A good improvement can be achieved by just using gender dependent models. In the previous section, we have seen that speakers can be separated out in acoustic space based on their age group with a reasonable accuracy. In this set of experiments, we explore how much increase in ASR accuracies can be obtained by exploiting the a priori knowledge about the gender and ages of the training and the test set speakers. Hierarchical acoustic models based on gender+age category are trained and used for decoding the test utterances. The experiments are carried out on the JNAS corpus due to the good balance of speakers in terms of age and gender.

The work by Baba et al. [2001] is similar in principle to this work. Based on the hypothesis that the acoustic space for elderly speakers is separable from that of younger speakers, separate speaker independent acoustic models are built for the two age groups. Those models were further adapted to each gender category to obtain significant improvements in performance. In our work on speaker clustering, we find that the top level clustering is principally based on the gender factor and hence we adopt the strategy of gender dependent models adapted to age category. Since the acoustic space for younger and elderly speakers is not completely separable, instead of training separate models for each category, we derive the gender-age based models by adaptation of speaker independent models thus allowing efficient sharing of data.

7.2.1 Experimental setup

The training data used for SI acoustic models in the baseline experiments on JNAS corpus (section 4.2.3) are balanced in terms of gender and age. Gender and age group specific models were built in a hierarchical structure from these models as shown in Figure 7.3. Using the MAP approach with τ 10, the speaker independent models are first adapted with the male and female subsets of the training set to get gender dependent models. These models are further adapted using 'gender + age group' specific subsets to create 4 more models. The mean, variance and mixture weight parameters are estimated for each model using the priors from the parent nodes.

The language models and the decoding setup are the same as in the setup described in section 4.2.3.

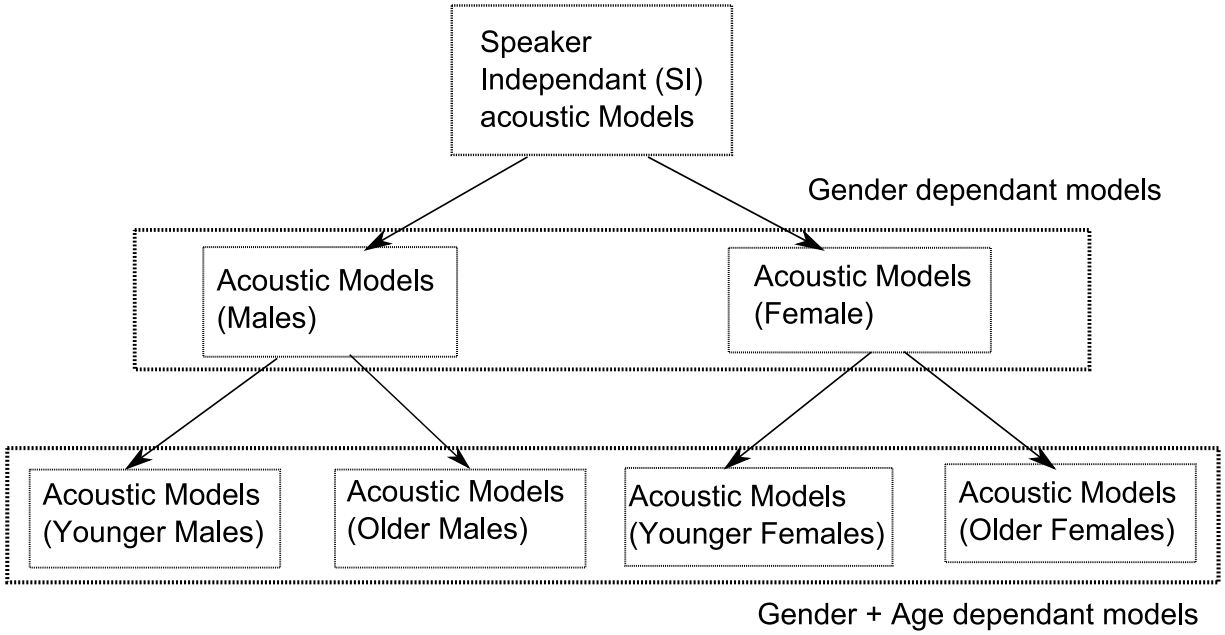


Figure 7.3: Training gender and age dependent acoustic models

7.2.2 Results

7.2.2.1 Gender dependent models

The test set utterances as described in the baseline experiments are decoded using the appropriate gender dependent model for each test speaker. We use the prior knowledge of the gender of the test speaker to choose the appropriate acoustic model. The WER results are tabulated in Table.7.2. Overall, there is an absolute improvement of 1.2% and 1.7% in WER over the baseline results (in section 4.2.3.6) for younger adults and older adults respectively. Older Males have the maximum improvements in WER of 2.5% as compared to the other groups.

	Younger adults	**Older adults**
All speakers	14.7	18.7
Male	15.1	21.4
Female	14.3	15.9

Table 7.2: Comparison of WERs (%) of younger and older adults in the JNAS corpus using gender dependant models

7.2.2.2 'Gender + Age' dependent models

Again the prior knowledge of the age of the test speaker is used to select one of the four acoustic models adapted to gender and age group. The results are tabulated in Table. 7.3. Consistent improvements in WERs over the baseline are observed for all groups. Interestingly, a further improvement over gender dependent models is obtained for older adults (both male and female speakers). From the results it appears that while it is beneficial for older adults to have the gender dependent models adapted to the older speaker set, it is not the case for younger adults. The performance is better for the younger adult group with a larger training pool (including speech from older speakers) of data.

	Younger adults	**Older adults**
All speakers	14.9	18.4
Male	15.7	21.1
Female	14.2	15.7

Table 7.3: Comparison of WERs (%) of younger and older adults using 'Gender + Age' dependant Models

7.3 Unsupervised hierarchical models

In the previous experiment, prior knowledge of the gender and age of the training and test set speakers was used to build the hierarchical models. This however is not feasible in practical systems. Most speech corpora are not annotated with gender and age details of the speakers.

7.3.1 Acoustic models

We use an unsupervised clustering approach described below to build the hierarchical models.

1. Using the speaker independent models, compute an MLLR transform for each speaker in the training set.

2. The distance between each pair of training speakers is computed using equation 7.2

3. The speakers are then clustered into groups using the repeated bisections method.

4. The SI models are adapted hierarchically using the data from the training speakers clustered at each node using MAP adaptation.

The training set age distribution in the hierarchical models is seen in Figure.7.4. The clustering at the first level turns out to be predominantly gender based. In the next level some pattern with age groups is observed. Nodes 1 and 2 are somewhat biased towards older males and females respectively, while nodes 0 and 3 are biased towards younger males and females respectively.

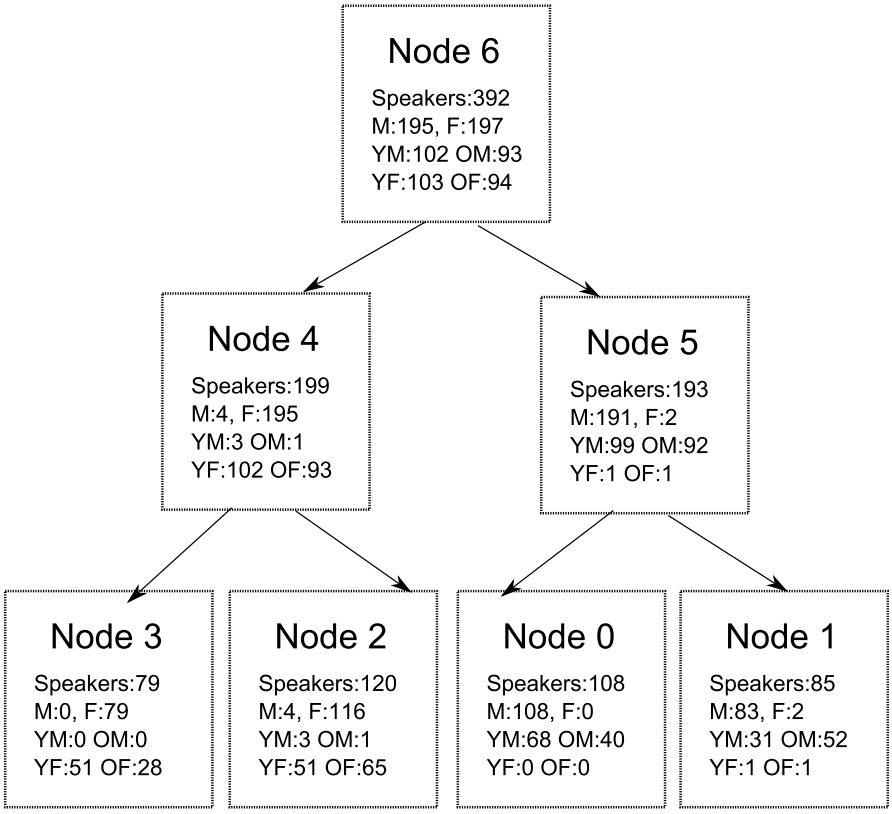


Figure 7.4: Unsupervised hierarchical models. Figure shows the age and gender statistics of training set speakers clustered at each node

7.3.2 Testing

Utterances from each test speaker are decoded against all the models in the hierarchical structure and the model that maximises the likelihood of the test set is chosen as the acoustic model for that speaker. However decoding with several model sets leads to undesirable increase in the computational time. To overcome this, we decode the first

test utterance from a speaker with all the models and select the model with the best likelihood score.

Typically, the models selected with the likelihood scores from the first utterance and scores accumulated over all the utterances are the same. There were a few cases where they were different, but in such cases the models were found to be parent and child nodes in the hierarchical tree and the choice of either of them did not have a significant impact on the ASR accuracies. Hence in all the results reported below, likelihood scores from the first utterance were used in model selection.

7.3.3 Results

The results with the best model choice for each test speaker are shown in Table. 7.4. It is seen that the results are comparable to the hierarchical models built using prior knowledge of age and gender.

	Younger adults	**Older adults**
All speakers	14.7	18.7
Male	15.0	21.3
Female	14.3	16.0

Table 7.4: WERs (%) of Younger and Older odults using unsupervised hierarchical models

The statistics for the number of test speakers choosing a particular model in the tree is shown in Figure 7.5. Female speakers do not show any pattern in picking age based models, while the male speakers seem to prefer models adapted to their age group. Interestingly, 11 out of the 15 older male speakers in the age group of 70-79 pick the model corresponding to Node 1 which has a higher proportion of older males in the training set.

7.4 Modifying HMM transition parameters

Often, there is a requirement to deploy ASR systems in environments where there is sufficient data available from younger speakers to build acoustic models, but zero resources available from older speakers. However the target set of users for the system might be older speakers, for instance in care homes.

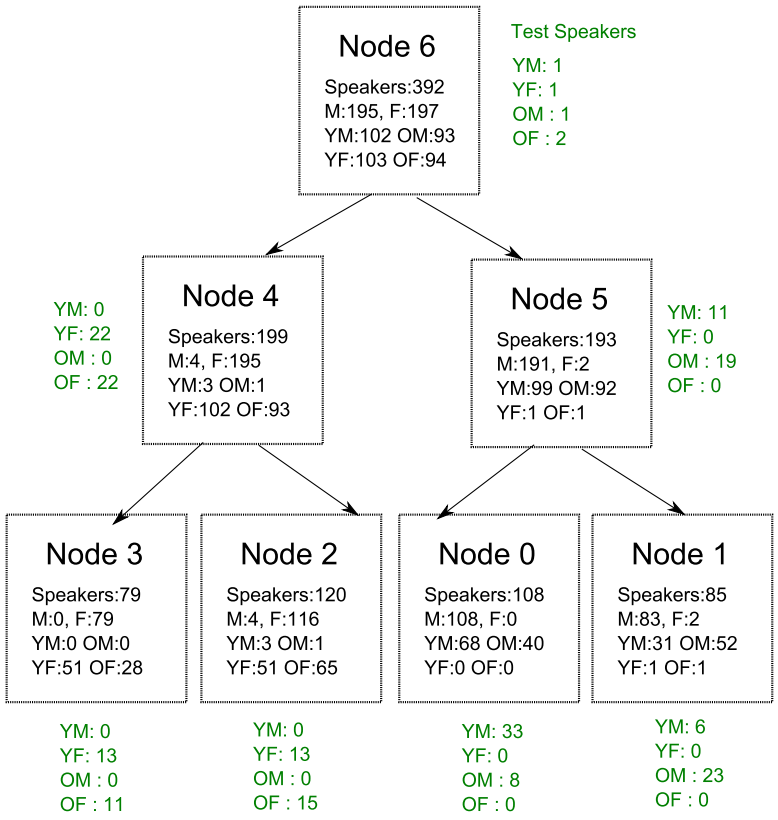


Figure 7.5: Unsupervised hierarchical model. At each node, the statistics of the test speakers selecting acoustic model corresponding to that node is shown.

Motivated by the results in section 6.3 where models trained on slower speech give much higher accuracies for older voices, we want to understand if it is possible to improve the ASR accuracies for older speakers by adjusting the state transition probabilities of the HMMs to increase state persistence and thereby model slower speech better.

Traditionally, hidden semi Markov models [Murphy, 2002] have been very popular models to capture variations in speaking rate. These models integrate the state duration probability distributions explicitly into the HMMs. In such models the unobservable state is semi-Markovian where the probability of state transition depends on the time duration elapsed since entry to the occupied state. Even though such models come with increased complexity, they have been shown to give significant improvements in ASR accuracies over the traditional HMMs [Oura et al., 2006].

In this section, we are interested in exploring the possibility of adaptation of HMM transition parameters to suit slower speech without any additional complexity in the model. Following procedure has been adapted for the transition parameter modification.

Let α be the desired fractional increase in the duration of occupancy of a state, $a_{i,i}$ be the probability of occupying the same state i and $a_{i,i\ 1}$ be the probability of transitioning to the next state in the following time instant. The modified parameter $\widehat{a}_{i,i}$ is related to the original parameter $a_{i,i}$ by the following relation.

$$\frac{1}{1-\widehat{a}_{i,i}} \quad \frac{1\ \alpha}{1-a_{i,i}} \tag{7.4}$$

Simplifying this relationship, and subtracting the residual weight added to $a_{i,i}$ from $a_{i,i\ 1}$, we get the following relations for the modified parameters.

$$\widehat{a}_{i,i} \qquad \frac{\alpha\ a_{\ ii}}{1\ \alpha} \tag{7.5}$$

$$\widehat{a}_{i,i\ 1} \qquad a_{\ i,i\ 1} - \frac{\alpha\ 1-a_{i,i}}{1\ \alpha} \tag{7.6}$$

7.4.1 Experimental results on the JNAS corpus

The experimental setup is similar to that described in section 4.2.3. The test set in this set of experiments is only the set of utterances from older adult speakers. The transition parameters of all the HMMs in the acoustic models are modified for α varying from 5% to 40% in steps of 5%. The WER results are shown in Table 7.5.

Word Error Rates (%)			
α	**Overall**	**Male**	**Female**
0	20.5	24.0	16.9
0.05	20.3	23.7	16.9
0.10	20.2	23.6	16.8
0.15	20.2	23.6	16.7
0.20	20.1	23.5	16.6
0.25	20.0	23.4	16.5
0.30	20.0	23.5	16.4
0.35	19.8	23.2	16.3
0.40	19.8	23.2	16.3

Table 7.5: WERs (%) on older speakers in the JNAS corpus using acoustic models with modified transition parameters

7.4.2 Experimental results on the SCOTUS corpus

The same experiment is repeated on the SCOTUS corpus. The test set is again only from older adult speakers. The results on SCOTUS corpus are displayed in Table. 7.6

7.4.3 Discussion

It is quite interesting to see some relative gains in accuracies on both the corpora. On JNAS corpus, the improvements in accuracies are quite reasonable, while on SCOTUS corpus, the improvements are minuscule. The possible cause is that the speaking rate for older speakers in JNAS corpus is substantially lower than younger speakers, while it is only marginally lower in SCOTUS corpus as seen in section 6.3.

If it is known apriori that the target users of the ASR system would be elderly speakers or people with a slower speaking rate, then there seems to be value in adjusting the transition parameters.

Although a value of 15% or 20% for α seems appropriate for older speakers, it is also not clear from the experimental results, what is the best choice for the value of α. This could possibly be better estimated in a maximum likelihood sense from a small development set from the target user.

Word Error Rates (%)			
α	**Overall**	**Male**	**Female**
0	40.4	38.8	46.1
0.05	40.3	38.8	46.0
0.10	40.3	38.8	45.9
0.15	40.3	38.8	45.9
0.20	40.4	38.9	45.9
0.25	40.4	38.9	45.8
0.30	40.4	38.9	45.8
0.35	40.9	39.3	46.4
0.40	40.8	39.3	46.4

Table 7.6: WERs (%) on older speakers in the SCOTUS Corpus with modified transition parameters

7.5 Summary

Though MFCC and PLP features are designed mainly to capture the phonetic characteristics in speech, they also capture meta information about speaker characteristics. Speaker clustering task into two age groups using these features achieves an accuracy of 70%. A method to compute the acoustic distance between two speakers using MLLR transforms is also introduced in the speaker clustering experiments.

Motivated by the separation of the speakers in acoustic space based on age and gender, the use of supervised hierarchical 'age and gender' models has been explored. Significant improvements in ASR accuracies are achieved using such models. It is also observed that for older adults these models outperform gender dependent models which is not the case for younger adults.

Using the speaker distance measure proposed, hierarchical models constructed in an unsupervised manner are also seen to give accuracies comparable to the supervised models.

Artificial modification of the transition parameters of the HMMs to cater for slower speaking rate of older speakers is explored. Favourable results are attained in this task on JNAS corpus.

Chapter 8

Speaker selection to augment adaptation data

In typical ASR based interactive voice response and spoken dialogue systems, only a few seconds of speech is generally available from a user to adapt the acoustic models to his/her voice. Linear regression based speaker adaptation techniques such as MLLR and MAPLR are widely used in such scenarios where the transformation matrices can be efficiently computed with a reasonable amount of data. However when the transforms are computed using a very small amount of adaptation data, the improvement in recognition accuracy using the adapted models can be low; indeed the accuracy with the adapted models can be lower than that with the speaker independent models.

To overcome this problem of sparse data, several approaches have been devised to characterise the test speaker and make better use of the data from the existing speakers. Eigenvoices [Kuhn et al., 2000] is one such idea where the test speaker is characterised as a linear combination of eigenvectors which are computed from speaker dependent (SD) models of the training set speakers. This approach however has limitations when applied to large vocabulary systems due to the need to generate several SD models and in the computation of speaker coefficients in the high dimensional Eigen-space.

Another approach to tackle data sparsity is to augment the adaptation data for the target speaker with speech data from other reference speakers acoustically close to the target speaker. The reference speakers can be a subset of the speakers used to train the SI models as well as other speakers whose data becomes available at a later stage. Such systems where more corpora becomes available for speaker selection can be easily envisaged in practical applications. In telephony based IVR systems, speech data can be collected as the system is used and the collected data can be made available

as a pool of reference speakers. In broadcast news, speech content is made available on daily basis from different speakers. Hence it makes sense in such scenarios, to build a speaker independent ASR system and use the data made available consequently, to improve the performance of the system.

Some related work based on this approach of speaker selection was conducted by Yoshizawa et al. [2005], where GMMs were trained for each reference speaker and the models that maximised the likelihood for the target speaker's adaptation data were chosen as the closest speakers. Wu and Chang [2001] built custom HMMs for each reference speaker using MLLR and the speakers whose models maximised the likelihood scores for forced alignment of the adaptation data were chosen as the reference speakers.

Recently, an approach to speaker recognition using MLLR transforms as feature vectors has been investigated [Stolcke et al., 2005, 2006]. The core idea is to concatenate the coefficients of the adaptation transforms into high dimensional vectors and use these vectors for speaker identification using SVM classifiers. Inspired by this work, we extend the idea of using transformation matrices as speaker features to identify the reference speakers acoustically closest to the target speaker. However, we do not use SVM classifiers since our task is different from speaker recognition. We use a distance metric based on transformations to compute the distance between speakers.

In this chapter we first explain the distance metric used and verify its validity on a speaker identification task. The speaker selection strategy to augment the adaptation data is then outlined. This is an interesting generic speaker adaptation strategy not specifically targeted for the older speakers. The experimental results on AMI corpus are discussed to illustrate the usefulness of this approach followed by the extension of the idea to SCOTUS corpus, where the results are analysed separately for younger and older speakers.

8.1 Distance measure

Stolcke et al. [2005] concatenate the coefficients of MLLR matrices to create high dimensional vectors and these vectors are used as speaker features. Such high dimensional vectors have been shown to have good discrimination properties for classification but the disadvantage of this approach is that it just treats the matrix as a vector and discards the property of the MLLR matrices that enable it to transform the means of HMMs to match the target speaker.

We propose a distance metric that takes advantage of the transformation defined by the MLLR matrices. Given MLLR matrices from two speakers, sample points in the acoustic space are transformed by the two matrices and the distance between the transformed points is calculated. We cluster the means of all the Gaussians in the SI model and choose the centroids of the partitions as the sample points. This ensures a good coverage over the acoustic space.

The distance d between two speakers whose MLLR transforms are represented by A_T and A_S is given by:

$$d \quad \sum_{k\ 1}^{K} \frac{A_T - A_S\, c_k}{c_k} \tag{8.1}$$

where c_k is the k^{th} mean of the K clusters computed from the means of all Gaussians in the speaker independent model.

This metric is in principle similar to matrix operator norm but instead of choosing the maximum, we sum over all the points being considered. It hence measures well the actual operation of the transformation matrix in the acoustic space.

8.2 Speaker identification task

In order to understand how well the proposed distance metric helps in identifying the closest transformation matrices, it was applied to a speaker identification task.

The experimental setup consisted of reference MLLR transforms and test MLLR transforms for a set of speakers. For each speaker, the utterances used in computing the reference and test transforms were disjoint. The task is to identify the closest reference transform for each test transform using the distance metric proposed and when the closest reference transform is from the same test speaker, it is treated as correct recognition.

The SCOTUS corpus was used for this task. Acoustic models comprised of 18 component GMMs per state. In all, the SI acoustic models comprised 59886 Gaussians over 3324 independent states.

A set of 100 speakers was used for the speaker identification task. To compute the reference and test MLLR transforms, about 40 seconds and 12 seconds of speech was used respectively for each speaker. A two class regression tree for speech and silence was used for the MLLR computation and only the speech transforms were used in distance computation.

The sample points in acoustic space to be used in calculating the distance, were selected as the centroids from k-means clustering on all the Gaussian means in the SI model. The task was repeated with various sizes of sample points viz., global mean, 100 clusters, 1000 clusters and using all the Gaussian Means in the SI model. A simple euclidean distance (Frobenius norm) between the co-efficients of the matrices was also used for comparison.

The results for this task are shown in Table 8.1. We observe that a set of 1000 points in the acoustic space is sufficient to achieve acceptable accuracy.

Distance measure	**Accuracy**
Euclidean Distance between transformation matrices	32%
Transformation of the global Mean of all Gaussians in SI model	36%
Transformation of 100 Cluster Means	97%
Transformation of 1000 Cluster Means	98%
Transformation of all the Gaussian Means	98%

Table 8.1: Speaker identification task

As mentioned earlier, the objective of this task is only to make a sanity check of the distance metric and hence the results have not been compared with other competing methods for speaker recognition.

8.3 Speaker selection

Given a set of N reference speakers, our task is to select a subset of these speakers who are acoustically closest to the target speaker T.

Denoting the transformation matrices for the target speaker as A_T, the i^{th} reference speaker as A_{R_i} $(i \quad 1 \ldots N)$ and using an identity matrix to represent the SI model $A_I \quad I_{\ m\ \ m} : 0_{m\ \ 1\ m \quad m\ \ 1}$,

Compute a linear transform A_T for the test speaker from the available adaptation data.

Compute the distances d_{TR_i} for $i \quad 1 \ldots N$ and $d_{\ TI}$.

Choose a subset of speakers satisfying $d_{TR_i} \leq d_{TI}$ to augment the adaptation data.

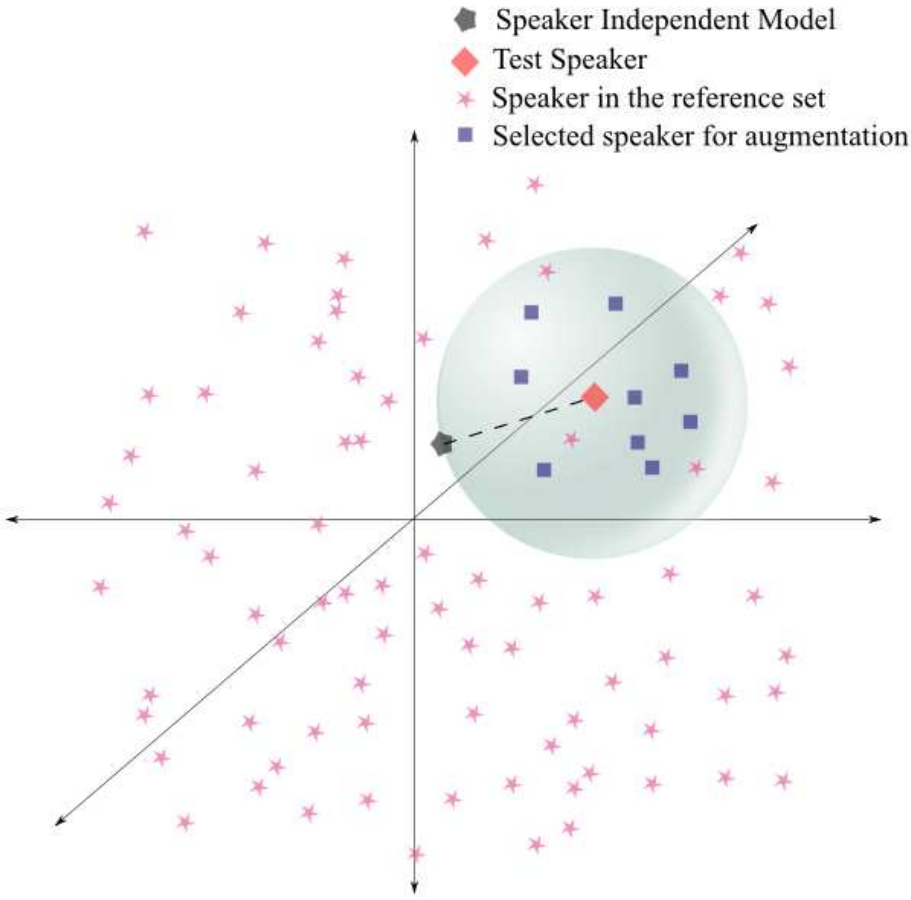


Figure 8.1: Speaker Selection

Recompute the linear transform for the test speaker using the augmented data.

To illustrate the speaker selection process, Figure. 8.1 shows the speakers in 3d space (generated using multidimensional scaling). The reference speakers selected for augmentation are the ones that lie within the spherical manifold with the target speaker at the center with a radius of d_{TI}. In practice, the dimension of the speaker space is large and the selection manifold is a high dimensional ellipsoid.

8.4 Experiments

The experiments were carried out on the AMI and SCOTUS corpora.

8.4.1 Experiments with the AMI corpus

8.4.1.1 Setup

Features

The waveforms were parametrised into 39 dimensional PLP based features with first and second order derivatives

Acoustic models

The ICSI-NIST-ISL acoustic models were MAP adapted with 40 hours of speech from the AMI corpus. With 8 and 16 Gaussian components per state for speech and silence respectively, the SI model comprised 3712 independent states with 29720 Gaussians in total.

Language models

Back-off bi-gram language models and vocabulary of size 50002 words were built using transcripts of several meeting corpora including Switchboard, Call Home, Fisher, ICSI, NIST, ISL and other web data resources [Hain et al., 2005b].

Reference set

The reference speaker set comprised of 69 speakers used to MAP adapt the SI models and 78 speakers not used in the training set. Each speaker had about 30 minutes of speech data on average.

Test set

The test speaker set in this corpus consisted of 42 speakers with 200 utterances as test data per speaker and a small adaptation set separate from the test set.

8.4.1.2 Procedure for speaker selection

The means of all the Gaussians in the SI acoustic models were clustered into 1000 groups using k-means clustering for each of the two corpora. The centroids of each of these clusters were used as the sample points for computing the acoustic distance between speakers. From each of the reference speakers' data, MLLR and MAPLR mean transforms were computed using a two class regression tree, one for speech and one for non-speech. Three sets of adaptation data were used with different amounts of data for the test speakers viz., 1)10-15 seconds of speech per speaker 2) About 30

seconds per speaker and 3) About 1 minute per speaker. Adaptation transforms were computed from all the adaptation sets using the actual transcripts for supervised case and using the hypotheses from first pass decoding for unsupervised case. For each of the test speakers, acoustically closest speakers were chosen as described in section 8.3.

8.4.1.3 Results

The baseline results for the AMI corpus are shown in Table 8.2. The baseline WER of 46.3% is much higher than 39% WER reported on the same corpus in [Hain et al., 2008], but our experimental setup is quite different from the original AMI setup. The ICSI-NIST-ISL models were adapted to the AMI domain with only 40 hours of in domain data, in order to keep aside seperate speaker set as reference speakers. Additionally the test set in our experimental setup is also different from the AMI set. Hence the difference in WER is not surprising.

The WERs with speaker adaptation are also tabulated in Table 8.2. When only 10-15 secs of adaptation data is available, it is seen that speaker adaptation is not optimal. The WERs increase in most cases.

Speaker independent	46.3		
Adaptation Data	***15s***	***30s***	***60s***
MLLR Supervised	50.2	46.0	43.9
MLLR Unsupervised	51.5	47.5	45.3
MAPLR Supervised	48.1	45.3	43.7
MAPLR Unsupervised	49.3	46.7	45.1

Table 8.2: AMI Corpus: Baseline results (WER %)

In order to understand if there is merit in adapting the acoustic models for a target speaker with speech from other speakers, an oracle style experiment was setup. For each of the 42 test speakers, the test utterances were decoded with MLLR transforms generated from each of the 78 reference speakers not present in the training set. The reference speakers were then sorted in increasing order of WER and the speech from top 10 speakers in the sorted list was chosen as the adaptation data. A WER of 45.4% was obtained using such adapted models. From the baseline results shown in Table 8.2, we observe that more than 30 seconds of adaptation data is required from the target speaker to achieve similar improvement in accuracy under the same experimental setup. This reinforces the hypothesis that speech data from acoustically similar

speakers can be shared to improve ASR WERs.

Reference Speakers	**Train set**			**Train + Add Set**		
Adaptation Data	***15s***	***30s***	***60s***	***15s***	***30s***	***60s***
MLLR Supervised	46.2	45.6	45.5	45.9	45.6	45.2
MLLR Unsupervised	47.4	46.2	45.8	46.8	45.9	45.4
MAPLR Supervised	46.1	45.7	45.4	45.8	45.7	45.3
MAPLR Unsupervised	47.1	46.1	45.7	46.2	45.9	45.7

Table 8.3: AMI Corpus: Results with augmented adaptation data (WER %)

Results with adaptation using augmented data using automatic speaker selection procedure are shown in Table. 8.3. It capture WERs using 1) only the training set speakers as reference speakers and 2) Training set speakers and additional speakers (Train + Add Set). The results show a significant reduction in WER with augmented adaptation data when the adaptation data is limited to 10-15 seconds. The WER reduction is significant at $p < 0.001$ using MAPSSWE test. As the adaptation data from the target speaker increases, the benefit from using other speakers' speech reduces.

Augmenting the adaptation data is seen to be particularly advantageous in the unsupervised case which is often the situation for practical systems. It is also observed that accuracies with MAPLR mean adaptation are overall better than MLLR mean adaptation. With augmented adaptation data, an improvement of 4.2% relative for supervised case and 4.5% relative for unsupervised case are achieved.

Using speakers additional to the training set speakers, a further improvement in recognition accuracies can be achieved as seen in Figure 8.2.

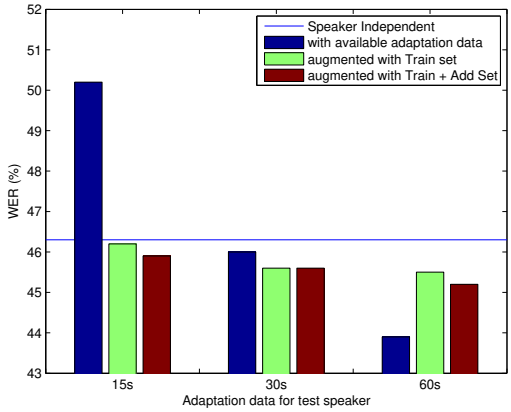
8.4.2 Experiments on SCOTUS Corpus

The main motivation to extend the experiment to SCOTUS corpus is to get a feel for the kind of improvements in ASR accuracies for younger and older adults using this approach.

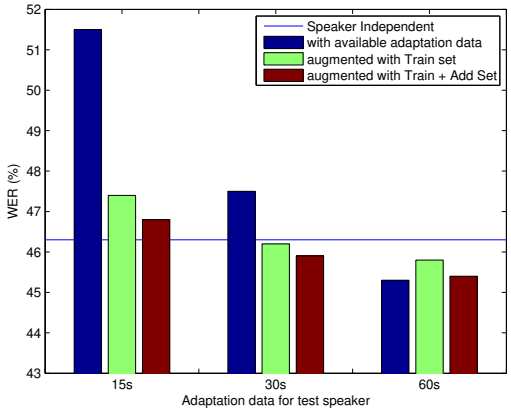
8.4.2.1 Setup

ASR system

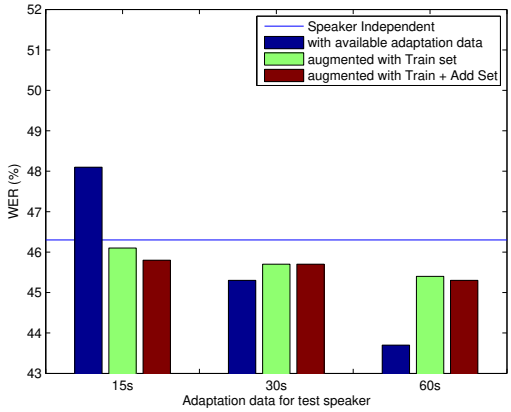
The SI acoustic models used are the same as those described in section 8.2. Back-off bigram language models and the vocabulary were constructed from the transcripts



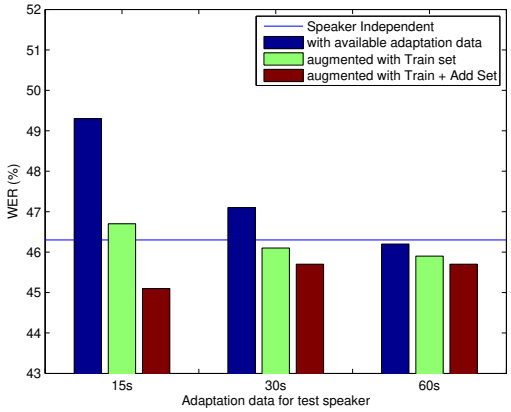
(a) MLLR Supervised



(b) MLLR Unupervised



(c) MAPLR Supervised



(d) MAPLR Unsupervised

Figure 8.2: Augmentation of adaptation data. Results on the AMI corpus

of the Supreme Court of the United States proceedings resulting in 23445 words types.

Reference Set

The reference speaker set consists of speech data from 267 training set speakers and 282 additional speakers not used in the training set. Each reference speaker had about 8 to 20 minutes of available data with an average of 12 minutes per speaker.

Test Set

The test speaker set comprised of 27 younger adults and 12 older adults disjoint from the training and additional speaker set. Each test speaker had about 60 minutes of data and a small set of about 3 minutes kept aside as the adaptation data.

8.4.3 Results

The same procedure as described in the AMI experiments above were followed even for the SCOTUS corpus.

The baseline results for the younger and older adults are shown in Tables 8.4 and 8.5 respectively. The results shown in these tables are with adaptation sets of 15 secs, 30 secs and 60 secs which are a subset of the 120 second adaptation data used for speaker adaptation results shown in section 4.2.1.1. It is again seen from these results, that speaker adaptation with very little adaptation data can be a tricky issue.

Speaker Independent	30.4		
Adaptation Data	***15s***	***30s***	***60s***
MLLR Supervised	30.9	30.3	29.8
MLLR Unsupervised	31.0	30.4	30.0
MAPLR Supervised	30.5	30.0	29.7
MAPLR Unsupervised	30.6	30.1	29.9

Table 8.4: SCOTUS Corpus: Baseline results (WER %) for *younger adult* speakers

The results with augmented adaptation data for younger and older adults are displayed in Tables 8.6 and 8.7 respectively.

From figures 8.3 and 8.4, it is seen that trends in WER improvements similar to those observed on AMI corpus are repeated on SCOTUS corpus as well. The gains for younger adults by data augmentation is higher than those for older adults. However, it must be noted that the number of older speakers in the training corpus as well as the additional set is extremely low. Hence suitable augmentation data for older speakers

Speaker Independent	40.4		
Adaptation Data	***15s***	***30s***	***60s***
MLLR Supervised	40.4	39.8	39.7
MLLR Unsupervised	41.0	40.3	40.0
MAPLR Supervised	39.8	39.5	39.3
MAPLR Unsupervised	40.3	39.8	39.8

Table 8.5: SCOTUS Corpus: Baseline results (WER %) for *older adult* speakers

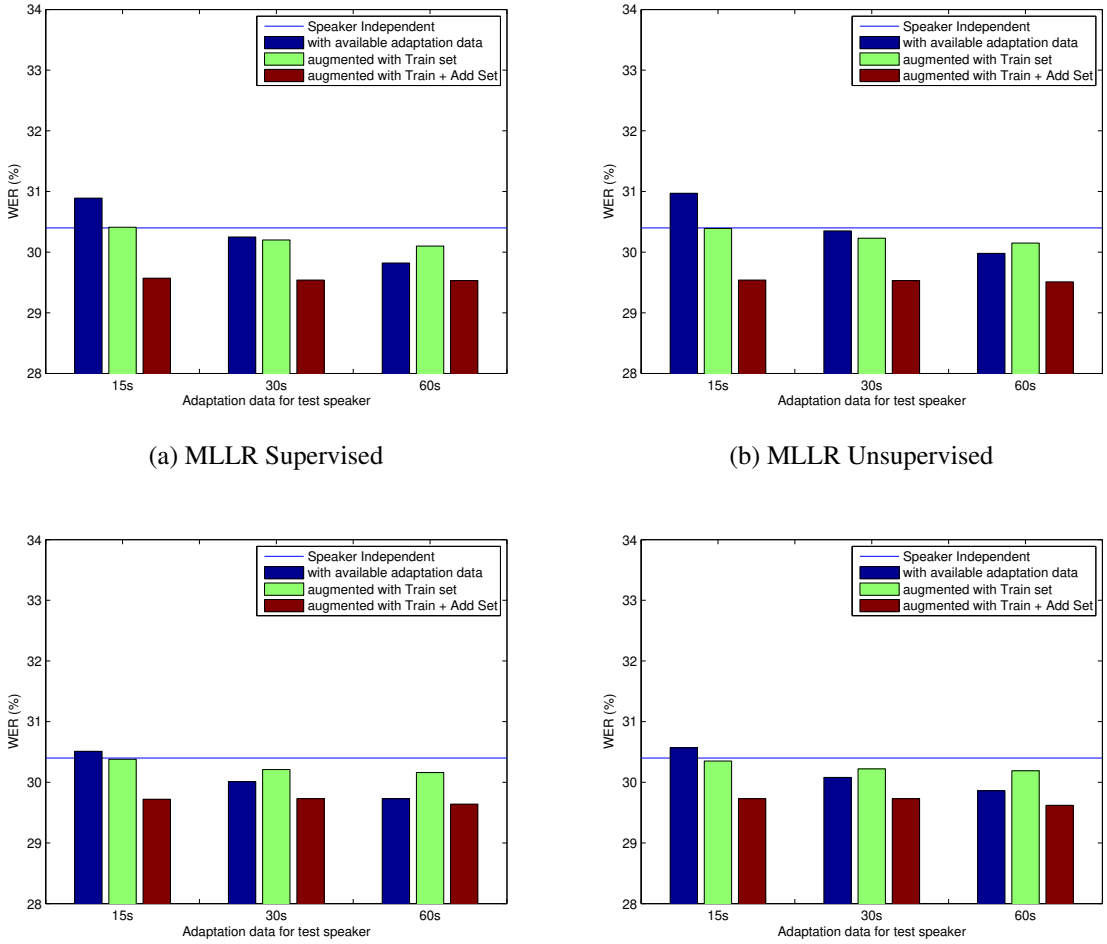
Reference Spkrs	**Train set**			**Train + Add set**		
Adaptation Data	***15s***	***30s***	***60s***	***15s***	***30s***	***60s***
MLLR Supervised	30.4	30.2	30.1	29.6	29.5	29.5
MLLR Unsupervised	30.4	30.2	30.2	29.5	29.5	29.5
MAPLR Supervised	30.4	30.2	30.2	29.7	29.7	29.6
MAPLR Unsupervised	30.4	30.2	30.2	29.7	29.7	29.6

Table 8.6: SCOTUS Corpus: Results with augmented adaptation data (WER %) on *younger adult* speakers

Reference Spkrs	**Train set**			**Train + Add set**		
Adaptation Data	***15s***	***30s***	***60s***	***15s***	***30s***	***60s***
MLLR Supervised	39.6	39.6	39.6	39.7	39.7	39.6
MLLR Unsupervised	39.6	39.6	39.7	39.7	39.7	39.7
MAPLR Supervised	39.7	39.7	39.7	39.7	39.7	39.5
MAPLR Unsupervised	39.6	39.7	39.7	39.8	39.7	39.7

Table 8.7: SCOTUS Corpus: Results with augmented adaptation data (WER %) on *older adult* speakers

is not readily available from the reference speakers. Despite that, there are minor improvements in WERs observed even for older adult speakers.



(a) MLLR Supervised

(b) MLLR Unsupervised

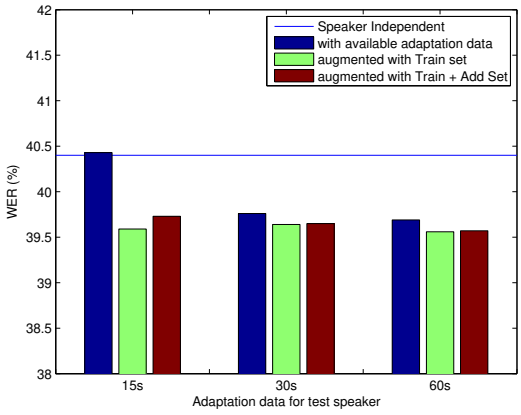
(c) MAPLR Supervised

(d) MAPLR Unsupervised

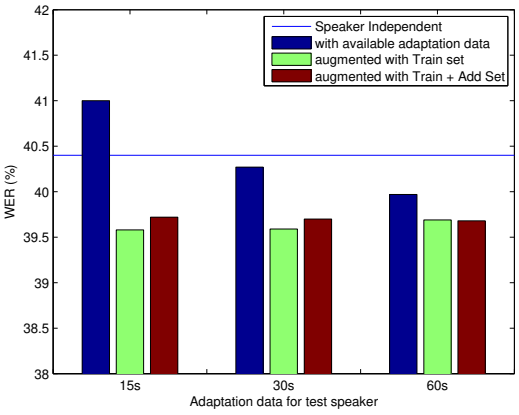
Figure 8.3: Augmentation of adaptation data. Results on the SCOTUS corpus for younger adult male speakers

8.4.4 Discussion

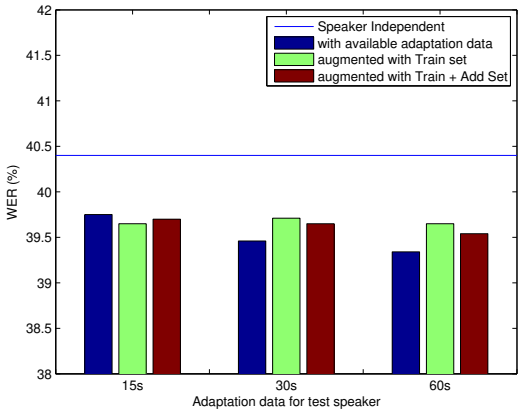
A simple and efficient method to improve the ASR accuracies with small amounts of adaptation data is described. Other approaches on similar tasks such as eigenvoices have been shown to improve performance in smaller systems, but scaling the eigenvoices approach as described in Kuhn et al. [2000] to our larger system led to Principal component analysis on a large matrix (2.5million x 250), which was computationally expensive. Due to the high dimensionality, all the eigenvectors generated had similar



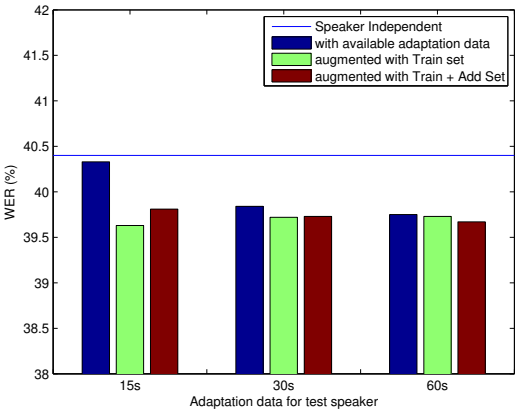
(a) MLLR Supervised



(b) MLLR Unsupervised



(c) MAPLR Supervised



(d) MAPLR Unsupervised

Figure 8.4: Augmentation of adaptation data. Results on the SCOTUS corpus for older adult male speakers

eigenvalues. Choosing the top 20 of them as basis results in loss of information and an increase in WER of 6.2%.

The distance metric proposed, is efficient in memory usage and computational complexity. The storage requirements are an $m \quad m \;\; 1$ matrix per reference speaker and K sample vectors in the acoustic space. The computation of the distance between two speakers involves only a few matrix operations. To speed up the computations $|c_k|$ terms could be precomputed and stored. To save on the time required for computing the regression transforms, sufficient statistics for the reference speakers can be computed offline.

Another feature of this approach is that there is no manual tuning or thresholding involved. The threshold is implicitly determined by the distance of the target speaker from the speaker independent model. If no reference speaker is close enough to the target speaker, only the available adaptation data for the test speaker is used. This approach is expected to work better with the availability of a larger and more varied reference speaker set in terms of age and gender. Furthermore, if the speech from a target speaker is available in the reference set, it is very likely to be selected first as augmentation data and improve the recognition accuracy significantly.

The MLLR/MAPLR WERs on AMI corpus with 15 seconds adaptation data are significantly higher as compared to the results with SI models. Despite this, the linear transform matrices still capture sufficient information about the speaker to be able to select augmentation data.

In both of our systems, the number of reference speakers were limited to a few hundred. If thousands of reference speakers are available in the selection pool, then computing the distance of the target speaker to all the speakers can be time consuming. A possible solution to this problem is

> project all the reference speakers (N) and SI model to a p dimensional space ($p << N$) using MDS.
>
> Select p non-coplanar speakers in this p dimensional space as reference points.
>
> For a target speaker compute the distance from these p reference speakers and project the target speaker into this reduced dimensional space using triangulation method.
>
> Select the augmentation speakers satisfying $\tilde{d}_{TR_i} \leq \tilde{d}_{TI}$, where $\tilde{d}$ is the euclidean distance.

8.5 Summary

In this chapter, a simple approach to compute distance between speakers using regression matrices as speaker features is proposed and discussed. Experimental results show that speakers acoustically close to the target speaker can be affectively selected from a pool of reference speakers to augment the adaptation data for the target speaker. It is a general speaker adaptation strategy that is applicable for speakers of all genders and ages. This approach works well when the adaptation data from the target speaker is very limited and gives significant reduction in WER. It is seen to be particularly useful when the adaptation is unsupervised which is often the case in practical deployments of ASR systems. However when sufficient adaptation data is available from the target speaker, augmenting it with speech from other speakers is not beneficial.

Results on SCOTUS corpus also suggest that for such a speaker selection strategy to work well for speakers of all age groups, a diverse set of reference speakers in terms of gender and age is desirable.

Chapter 9

Conclusions

9.1 Summary

In this thesis, we focused on the problem of Automatic Speech Recognition for the domain of older voices. While there are several subsystems in an ASR system, the problem was investigated from an acoustic modeling point of view. The main questions that this work attempted to answer are as follows:

> What is the impact of changes that take place in speech production due to ageing on ASR accuracies?
>
> How can the acoustic modeling component of the ASR system be improved for the target domain of older speakers?

In order to answer the above questions, several research objectives as mentioned in section 1.2 were set a forth. The outcomes of the experiments to address each of those objectives are summarised below.

ASR accuracies for older voices

To start with, baseline experiments were set up with three different corpora viz., SCOTUS, MATCH and JNAS having substantial amount of speech from older speakers. Results on SCOTUS and MATCH corpora show about 9-11% higher WERs for older adults as compared to younger adults which are consistent with results from such previous studies by Wilpon and Jacobsen [1996] and Anderson et al. [1999] on different corpora. Longitudinal study on the WERs of older speakers in the SCOTUS corpus

showed increase in WER with age. The differences in the WER could not be alleviated even with the use of state-of-the-art speaker adaptation and speaker normlisation techniques. Results on JNAS corpus which has an extensive set of speakers from both the age groups display about 4.5% higher WER for older speakers. The performance deterioration was found to higher for elderly males with an increase of 7.7% in WER. These results laid the foundation for further investigation into the possible causes and look for other strategies to improve the acoustic models.

Impact of changes in glottal source characteristics on ASR accuracies

The investigation was started with careful analysis of glottal source characteristics. The following parameters that correspond to glottal source characteristics were analysed : Fundamental frequency, jitter, shimmer and Harmonic to noise ratio. These parameters strongly correlate with the change in pitch, hoarseness and breathiness usually associated with older voices.

Comparative analysis on the male speakers on SCOTUS corpus showed a decrease in fundamental frequency measures and an increase in jitter and shimmer measures that are associated with vocal cord instability. Harmonic to noise ratio analysis showed little differences between the two age groups. The parameters where there was significant change with ageing, were then then analysed carefully to understand the impact of the changes on ASR accuracies. Speech from younger adults was artificially modified to reflect the changes observed in the above parameters and ASR accuracies compared. Decrease in $F0$ by 10% increased the WER by 1.1% absolute. However, it was shown that this can be compensated to some extent using vocal tract length normalisation. ASR experiments with artificial increase in jitter and shimmer measures showed that these changes do not have a significant impact on WERs.

While Glottal source parameters provide strong acoustic cues of ageing and help in perception of speaker age by humans as well as machines, it is interesting to find that their impact on ASR accuracies in minimal.

Study of the articulatory changes in older voices

One important characteristic that is often associated with older voices is less precise articulation. An ASR system was set up in a phoneme loop decoder mode for computing phoneme recognition accuracies. Phoneme errors for younger and older adults on two different corpora viz., SCOTUS and JNAS were compared. The motivation

behind this set of experiments was to see if any strong patterns emerge in the changes in articulatory pattern of certain phonemes with ageing. Although the results show that some lower vowels are commonly more affected due to ageing, there is a great degree of variability in the results across different corpora and across different speakers. Analysis of vowel space area bounded by the first and second formants indicated vowel centralisation with ageing thereby decreasing the discrimination capacity between vowels.

Impact of slower speaking rate on ASR accuracies

Another important articulatory change associated with older voices is the decrease in speaking rate. While there has been some research work previously [Fosler-Lussier and Morgan, 1999] that concludes that WERs for faster speech is higher, there has not been much insight into the impact of slower speech. To understand this better, transition parameters of the acoustic models trained on older speakers were modified to reflect the slower speech of older speakers and vice versa. Experimental results on ASR WERs showed that there is an increase in insertion errors with slower speech decoded with acoustic models trained on faster speech and this impact was found to be more pronounced for older male speakers than their female counterparts.

Hierarchical models based on gender and age group

With the results from the analysis of glottal source characteristics suggesting little impact on ASR accuracies and with high variability of the results in articulatory changes across corpora and speakers, it became evident that it is difficult to exploit such information directly to improve the ASR accuracies.

Hence in order to answer the problem of better acoustic models for older speakers, we first looked at experiments to understand the acoustic separability of speakers in terms of their age group based on the feature vectors used in ASR systems. The speaker age group recognition task gave a classification accuracy of around 70% using two different approaches based on SVMs and repeated bisection clustering.

Inspired by these results, we looked at the use of hierarchical acoustic models built in supervised manner based on gender and age groups of the training set speakers. Results on JNAS corpus showed a 2% absolute improvement in accuracies over the baseline results. The results also indicated that there is additional gain in accuracies for older adults in using gender and age dependent models over just gender dependent

models. Hierarchical models built in an unsupervised manner were also seen to achieve comparable accuracies to those of supervised models.

We then looked at a simple strategy to modify the transition parameters of the HMMs such that there is an increase in state persistence to suit slower speaking rate. Results indicated that there is value in such a method especially when there is a mismatch in the speaking rate of training and test set speakers.

Augmentation of adaptation data with speech from other acoustically close speakers

Typically the adaptation data available for a target speaker in practical deployments of ASR systems is quite limited. To address this problem, a method to use speech data from acoustically close speakers for adaptation was explored.

A distance metric based on the adaptation transforms was proposed to compute the acoustic distance between speakers. Using this metric, A strategy was devised to select speech data from acoustically close speakers to augment the adaptation data for the target speaker. Adaptation with augmented data was seen to give significant improvements in accuracies especially when the adaptation data is limited to a few seconds. The improvements in accuracies were found to be even better in unsupervised adaptation. While the method is a general purpose technique for all speakers, the improvements for older speakers was analysed in particular on the SCOTUS corpus. Favourable results were achieved for both younger and older adults on this corpora.

9.2 Future work

The three corpora used in this thesis have certain limitations. The SCOTUS corpus has few older speakers and is very specific to the legal domain where the speaking styles of the speakers are constrained. MATCH corpus is conversational in nature but the amount of speech data available is limited. JNAS has a good balance of speakers in age and gender but it is read speech. For this kind of study, it is desirable to have large amount of conversational speech with a good balance of speakers in terms of age and gender.

An interesting finding that has come out from the experiments on MATCH corpus is the difference in language patterns used by younger and older adults while interacting with dialogue systems. While the focus of this thesis has been on acoustic modeling,

it would be interesting to carry forward the work to understand the subtleties involved in the language modeling for older adults especially in conversational speech.

The age range of the older speakers used in our experiments are between 60-80 years with the bulk of them in the age group of 60-70 years. It would be interesting to extend the experiments on speaker characterisation in acoustic space for speakers further into older age. Speakers in age range above 70 years of age is of particular interest in this research since they are the real target group that would benefit from research in this direction. Data from such older speakers would also highlight other disfluencies such as breathiness and slurred speech due to vocal ageing that are not prominent in the data used in this thesis.

In this thesis, the methodologies employed are targeted towards improving the ASR systems for the domain of older voices. However what is interesting is to build upon the findings in this thesis to construct more generalised acoustic models that are agnostic to variations in age.

Appendix A

Appendix: Experimental result tables

Word Error Rate (WER) %				
	Younger adult voices	**Older adult voices**	**Difference**	**p-value**
Overall	30.4	40.4	10.0	< 0.001
Male	30.1	38.8	8.7	< 0.001
Female	32.4	46.1	13.7	< 0.001

Table A.1: Comparison of WER (%) on *younger adult* and *older adult* voices in the SCOTUS corpus

Word Error Rate (WER) %				
	Younger adult voices	**Older adult voices**	**Difference**	**p-value**
Overall	29.6	38.7	9.1	< 0.001
Male	29.5	38.1	8.6	< 0.001
Female	30.0	41.0	11.0	< 0.001

Table A.2: Comparison of WER (%) using MLLR speaker adaptation on *younger adult* and *older adult* voices in the SCOTUS corpus

Word Error Rate (WER) %				
	Younger adult voices	**Older adult voices**	**Difference**	**p-value**
Overall	28.7	38.6	9.9	< 0.001
Male	28.7	37.9	9.2	< 0.001
Female	28.2	41.3	13.1	< 0.001

Table A.3: Comparison of WER (%) using vocal tract length normalisation on *younger adult* and *older adult* voices in the SCOTUS corpus

Word Error Rate (WER) %				
	Younger adult voices	**Older adult voices**	**Difference**	**p-value**
Overall	27.9	37.6	9.7	< 0.001
Male	27.9	37.1	9.2	< 0.001
Female	28.1	39.4	11.3	< 0.001

Table A.4: Comparison of WER (%) using speaker adaptive training on *younger adult* and *older adult* voices in the SCOTUS corpus

Word Error Rate (WER) %							
	Speaker ID						
Age	**02**	**03**	**04**	**05**	**07**	**08**	**10**
59				32.0			
60				34.9		32.4	
61				33.0		32.7	
62			41.4	31.1		34.5	
63		34.7	42.8	32.0		33.6	
64		35.8	44.3	31.2		33.6	
65		34.2	45.0	31.9		35.1	
66		38.3	43.1	32.2	49.8	33.7	
67		36.4	47.0	33.8	51.1	35.0	
68		40.3	45.0		53.7	35.7	
69		35.5	43.5		56.2		41.0
70		37.4	47.7		55.3		42.6
71		38.9			60.4		41.2
72					55.3		44.2
73					57.7		45.3
74					61.8		42.7
75							44.0
79	42.6						
80	44.2						
81	43.4						
82	46.2						
83	40.0						
84	43.6						
85	44.7						
86	45.4						
87	50.6						

Table A.5: WER (%) with increasing age on older adult voices in the SCOTUS corpus

Word Error Rate (WER) %							
	Speaker ID						
Age	02	03	04	05	07	08	10
59				31.3			
60				32.5		31.7	
61				31.1		31.5	
62			40.4	29.4		32.7	
63		33.7	40.4	30.0		31.8	
64		34.3	42.9	29.1		31.6	
65		32.4	43.6	30.5		33.8	
66		36.3	41.4	30.6	40.0	32.7	
67		36.1	45.1	32.5	40.7	33.6	
68		37.8	43.6		41.3	34.7	
69		34.0	42.6		46.9		35.0
70		35.1	46.6		43.1		36.5
71		36.6			45.6		35.3
72					43.2		38.3
73					44.5		37.6
74					47.6		36.1
75							37.6
79	40.1						
80	42.1						
81	41.2						
82	44.1						
83	38.7						
84	41.3						
85	41.4						
86	41.3						
87	47.0						

Table A.6: WER (%) with increasing age on older adult voices using MLLR speaker adaptation in the SCOTUS corpus

Word Error Rate (WER) %			
Young speakers		Older speakers	
Language model	WER	Language model	WER
LM-All-1	24.0	*LM-All-1*	39.3
LM-Young-1	22.0	*LM-Young*	45.6
LM-Older	25.9	*LM-Older-1*	40.4

Table A.7: Comparison of WER (%) of young and older voices on MATCH corpus using different language models

Word Error Rate (WER) %			
Young speakers		**Older speakers**	
Acoustic model	**WER**	**Acoustic model**	**WER**
Baseline (AMI)	22.4	*Baseline (AMI)*	33.9
AMI + MATCH Young-1	10.8	*AMI + MATCH Young*	38.3
AMI + MATCH Older	13.8	*AMI + MATCH Older-1*	25.2

Table A.8: Comparison of WER (%) of younger and older voices on MATCH corpus using different acoustic models.

Phoneme	**Younger adult males**		**Older adult males**	
	F1 (Hz)	**F2 (Hz)**	**F1 (Hz)**	**F2 (Hz)**
aa	725.1	1395.0	671.4	1409.9
uw	429.6	1761.1	404.7	1786.7
iy	429.0	2026.5	378.7	2054.9
ae	622.8	1727.2	567.8	1680.0
ih	481.6	1948.6	507.7	1976.5
ax	519.4	1825.2	556.3	1916.1
eh	571.1	1677.7	518.9	1700.9
er	552.2	1605.4	491.6	1653.0
ah	582.7	1578.7	558.1	1659.7
ao	579.6	1293.6	579.7	1327.5
uh	477.8	1610.2	430.6	1661.4

Table A.9: F1 and F2 for the monophthongs in the SCOTUS corpus

Correct recognition (%) of phonemes				
Phoneme	**Occurrence (%)**	**Younger male adults**	**Older male adults**	**Difference**
aa	1.8	61.4	46.7	14.7
ae	3.6	46.6	36.8	9.8
ah	2.2	50.2	52.8	-2.6
ao	1.6	59.6	48.9	10.7
aw	0.4	67.8	52.2	15.6
ax	9.0	39.5	40.9	-1.4
ay	1.5	72.6	69.9	2.7
b	1.6	62.1	62.5	-0.4
ch	0.5	71.4	69.4	2.0
d	3.9	37.2	41.2	-3.9
dh	3.9	53.6	53.8	-0.2
eh	2.9	48.4	48.6	-0.2
er	2.2	63.5	53.2	10.3
ey	1.7	75.3	74.6	0.8
f	1.6	80.2	75.5	4.6
g	0.8	60.4	67.8	-7.3
hh	0.9	60.8	49.2	11.6
ih	6.5	44.6	44.1	0.6
iy	2.9	69.5	67.4	2.1
jh	0.7	69.0	64.2	4.8
k	3.6	63.3	64.9	-1.6
l	3.2	59.4	55.5	3.9
m	2.2	71.9	68.7	3.2
n	7.2	57.5	54.8	2.7
ng	0.9	71.1	66.6	4.5
ow	1.1	63.1	64.8	-1.7
oy	0.1	80.2	78.3	1.9
p	1.9	68.8	67.2	1.6
r	4.4	55.5	53.9	1.6
s	5.4	75.2	75.3	-0.1
sh	0.9	77.3	80.1	-2.8
t	8.7	33.4	31.1	2.3
th	0.6	48.3	44.1	4.3
uh	0.5	65.2	68.3	-3.2
uw	1.3	63.1	63.3	-0.2
v	1.9	58.4	57.5	0.9
w	2.1	74.2	71.7	2.5
y	1.0	68.2	69.6	-1.5
z	2.7	68.1	64.2	3.9
zh	0.1	73.9	74.7	-0.7

Table A.10: Correct recognition (%) of phonemes on younger and older adult males in the SCOTUS corpus

Correct recognition (%) of phonemes				
Phoneme	**Occurrence (%)**	**Younger male adults**	**Older male adults**	**Difference**
a	13.9	71.1	65.1	6.0
i	9.4	65.0	52.6	12.4
u	5.7	61.4	57.8	3.6
e	6.2	68.2	62.0	6.2
o	9.7	68.4	71.3	-2.9
a:	0.1	85.6	82.7	2.9
i:	0.2	83.3	78.8	4.5
u:	0.8	34.0	87.8	-53.8
e:	0.8	81.6	86.3	-4.7
o:	2.6	89.2	88.0	1.2
N	4.0	83.1	80.8	2.3
w	1.5	88.3	85.1	3.2
y	1.3	67.9	55.0	12.9
j	1.4	78.8	76.8	2.0
ky	0.5	85.4	90.8	-5.4
by	0.0	52.2	83.3	-31.1
gy	0.1	68.7	47.6	21.1
ny	0.1	100.0	88.9	11.1
hy	0.1	86.1	62.2	23.9
ry	0.2	75.0	64.8	10.2
py	0.0	78.6	100.0	-21.4
p	0.4	76.8	78.9	-2.1
t	4.8	80.2	78.2	2.0
k	6.8	72.9	71.1	1.8
ts	1.0	77.7	70.3	7.4
ch	1.0	77.2	78.9	-1.7
b	1.1	81.2	74.8	6.4
d	2.4	73.4	73.2	0.2
g	2.4	55.4	49.5	5.9
z	0.6	81.8	64.3	17.5
m	2.7	80.1	70.4	9.7
n	5.3	59.0	60.6	-1.6
s	3.1	87.3	76.1	11.2
sh	2.6	78.4	79.7	-1.3
h	1.4	78.3	80.7	-2.4
f	0.4	90.2	78.3	11.9
r	4.0	73.1	63.1	10.0
q	1.3	85.5	87.9	-2.4

Table A.11: Correct recognition (%) of phonemes on younger and older adult males in the JNAS corpus [101]

[101] Phonemes *my* and *dy* have not been shown in the table since no instances were found in the test set for atleast one of the age groups

Duration (Frames) per phoneme			
Phoneme	**Occurrence**	**Younger adult males**	**Older adult males**
aa	1.8	9.6	10.0
ae	3.6	8.9	9.6
ah	2.2	6.9	7.6
ao	1.6	9.0	10.3
aw	0.4	13.0	14.0
ax	9.0	6.1	7.1
ay	1.5	11.5	13.3
b	1.6	7.5	8.3
ch	0.5	10.3	10.9
d	3.9	7.0	8.2
dh	3.9	6.0	6.9
eh	2.9	7.3	8.0
er	2.2	10.0	11.0
ey	1.7	11.0	12.3
f	1.6	11.1	12.2
g	0.8	6.2	7.2
hh	0.9	8.0	7.6
ih	6.5	6.4	7.5
iy	2.9	10.4	11.7
jh	0.7	8.8	9.3
k	3.6	8.7	9.3
l	3.2	8.1	9.1
m	2.2	6.8	7.6
n	7.2	7.3	8.5
ng	0.9	8.5	10.3
ow	1.1	11.5	12.7
oy	0.1	10.5	11.9
p	1.9	9.3	10.5
r	4.4	6.5	7.5
s	5.4	10.1	10.5
sh	0.9	9.0	8.8
t	8.7	8.4	9.7
th	0.6	8.6	9.6
uh	0.5	5.1	5.8
uw	1.3	10.6	12.0
v	1.9	7.4	8.7
w	2.1	6.7	7.6
y	1.0	6.3	6.9
z	2.7	11.6	12.7
zh	0.1	7.8	8.2

Table A.12: Speaking Rate (Frames/Phoneme) on the SCOTUS Corpus

Phoneme	Occurrence	**Expected Frames/Phoneme** (using model parameters)				**Frames/Phoneme** (using forced alignment)			
		Male		**Female**		**Male**		**Female**	
		Young	**Old**	**Young**	**Old**	**Young**	**Old**	**Young**	**Old**
a	13.87	6.8	8.5	7.5	8.5	6.7	8.4	7.6	8.4
a:	0.15	13.2	18.1	13.9	18.3	8.3	9.7	8.9	10.0
b	1.11	6.4	8.4	7.1	8.9	6.7	8.0	7.4	8.4
by	0.04	11.0	14.0	10.8	14.0	8.8	9.5	8.1	10.6
ch	1.01	11.4	13.8	12.2	14.9	10.5	11.9	10.8	12.0
d	2.40	4.7	6.9	5.1	7.1	5.1	7.1	5.6	7.3
dy	0.00	14.8	17.3	13.0	22.5	12.3	14.4	11.5	10.8
e	6.17	7.3	9.5	8.2	9.8	7.6	9.4	8.4	9.7
e:	0.76	13.7	18.5	15.0	19.6	10.5	12.6	11.1	13.3
f	0.41	8.4	12.3	9.2	12.3	7.9	11.1	7.7	11.8
g	2.45	4.9	6.4	5.3	7.0	4.8	6.0	5.3	6.9
gy	0.11	10.9	13.1	10.7	12.8	8.8	9.5	8.4	9.4
h	1.41	7.4	9.7	8.7	10.3	7.3	8.5	8.4	9.3
hy	0.13	9.8	11.8	11.0	12.7	8.3	9.1	9.2	10.6
i	9.36	6.7	8.8	7.1	8.7	7.1	9.2	7.5	8.6
i:	0.21	14.5	19.9	16.0	20.0	9.4	12.0	10.1	11.9
j	1.43	9.3	11.8	9.6	11.8	9.3	11.0	9.7	11.0
k	6.77	7.1	8.6	7.8	10.1	6.9	7.7	7.1	8.8
ky	0.47	11.1	14.3	12.0	15.5	8.7	10.0	9.0	10.8
m	2.70	7.0	8.9	7.8	9.8	7.8	9.3	8.5	10.2
my	0.01	11.6	14.2	11.9	15.3	9.0	8.9	7.7	11.2
N	4.02	7.6	11.7	7.7	11.3	7.4	11.6	7.2	10.9
n	5.30	4.7	6.9	5.6	7.8	5.1	7.4	5.9	8.5
ny	0.07	12.5	14.4	13.4	15.5	10.0	11.6	11.1	12.3
o	9.66	6.4	9.0	7.1	8.7	6.8	9.6	7.6	8.8
o:	2.55	13.0	18.0	14.4	19.0	12.3	15.2	13.3	16.2
p	0.38	7.1	9.3	7.7	11.5	5.1	6.3	5.1	9.8
py	0.02	9.8	11.8	9.5	12.7	6.2	6.7	5.8	7.8
q	1.29	9.1	13.5	9.8	14.1	10.2	15.0	11.4	15.0
r	4.03	4.7	6.0	5.2	6.3	5.3	6.5	5.7	6.8
ry	0.21	9.3	11.9	9.8	12.2	7.6	9.2	8.2	9.7
s	3.09	10.5	13.7	11.2	14.6	10.7	12.7	11.0	13.2
sh	2.62	11.6	14.8	12.8	16.0	12.1	14.4	12.9	15.2
t	4.82	5.7	7.9	6.5	9.4	5.5	6.8	5.8	7.9
ts	1.00	11.4	13.9	12.8	15.1	10.7	12.0	11.5	12.8
u	5.65	5.5	6.5	6.1	6.8	5.3	6.4	6.0	6.7
u:	0.83	10.8	15.8	12.1	17.0	9.2	12.7	9.9	13.9
w	1.54	8.1	10.3	8.7	10.4	8.3	9.6	9.0	10.2
y	1.29	7.5	10.0	8.3	9.8	7.2	9.4	8.3	8.9
z	0.64	8.2	10.3	8.5	10.4	7.7	8.5	8.3	8.9

Table A.13: Speaking rate (Frames/Phoneme) on the JNAS corpus

Bibliography

Jitendra Ajmera and Felix Burkhardt. Age and Gender Classification using Modulation Cepstrum. In *Odyssey*, 2008. 59

T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *International Conference on Speech and Langauge Processing*, volume 2, pages 1137–1140, 1996. 52

Stephen Anderson, Natalie Liberman, Erica Bernstein, Stephen Foster, Erin Cate, Brenda Levin, and Randy Hudson. Recognition of elderly speech and voice-driven document retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 145–148, 1999. 2, 61, 134

S. Austin, R. Schwartz, and P. Placeway. The forward-backward search algorithm. In *International conference on Acoustics, Speech and Signal Processing*, pages 697–700, 1991. 45

A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano. Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2 (Electronics).(USA)*, 87(7):49–57, 2004. 2, 61

Akira Baba, Shinichi Yoshizawa, Miichi Yamada, Akinobu Lee, and Kiyohora Shikano. Elderly acoustic model for large vocabulary continuous speech recognition. In *Eurospeech*, pages 1657–1660, 2001. 110

Amil C. Bach, Francis L. Lederer, and Robert Dinolt. Senile changes in the laryngeal musculature. *Archives of Otolaryngology*, 34:47–56, 1941. 9

Lars Bäckman, Brent J. Small, and Ake Wahlin. *Handbook of the psychology of aging*, chapter Aging and memory: Cognitive and biological perspectives, pages 349–377. Academic Press, San Diego, CA, US, 2001. 2

L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Context dependent modeling of phones in continuous speech using decision trees. In *Proceedings of DARPA Speech and Natura Language Processing Workshop*, pages 264–270, 1991. 37

J. K. Baker. The Dragon system – an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975. 28

Kristin K. Baker, Lorraine Olson Ramig, Erich S. Luschei, and Marshall E. Smith. Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and aging. *Neurology*, 51:1592–1598, 1998. 9

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33:5–22, 2001. 65

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970. 34

Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of mathematical statistics*, 37(6):1554–1563, 1966. 28

LF Black and RE Hyatt. Maximal respiratory pressures: Normal values and relationship to age and sex. *The American review of respiratory disease*, 99:696–702, 1969. 7

FR Bode, J Dosman, RR Martin, H Ghezzo, and PT Macklem. Age and sex differences in lung elasticity, and in closing capacity in nonsmokers. *Journal of applied physiology*, 41:129–135, 1976. 7

Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, pages 97–110, 1993. 15

Paul Boersma. Praat: A system for doing phonetics by computer. *Glot International*, 5:9/10:341–345, 2001. 13, 81

B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pages 209–243, 1963. 21

Thorsten Brants and Alex Franz. Web 1t 5-gram version 1, linguistic data consortium, philadelphia, 2006. 39

J.S. Bridle and M.D. Brown. A data adaptive frame rate technique and its use in automatic speech recognition. In *In proceedings of the Institute of Acoustics Autumn Conference*, 1982. 21

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L.Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992. 42

Markus Bruckl. Women's vocal aging: A longitudinal approach. *Interspeech, Antwerp, Belgium*, pages 1170–1173, 2007. 15

Markus Bruckl and Walter Sendlmeier. Aging female voices: An Acoustic and Perceptive Analysis. *Proc of ISCA workshop Voqual '03*, pages 163–168, 2003. 13, 15, 17

Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. 65

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 106

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318, 1996. 39, 41

Cristina Chesta, Olivier Siohan, and Chin-Hui Lee. Maximum a posteriori linear regression for hidden Markov model adaptation. In *Eurospeech*, pages 211–214, 1999. 56

Jordan Cohen, Terri Kamm, and Andreas G. Andreou. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *Journal of Acoustic Society of America*, 97(5):3246–3247, 1995. 47

H.C. Crow and J.A. Ship. Tongue strength and endurance in different aged individuals. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 51: M247–M250, 1996. 11

Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980. 21, 59

Rogrio Aparecido Dedivitis, Mrcio Abraho, Manoel de Jesus Simes, Osvaldo Alves Mora, and Onivaldo Cervantes. Cricoarytenoid joint: histological changes during aging. *Sao Paulo Medical Journal*, 119:89–90, 2001. 8

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical society*, 39:1–38, 1977. 28, 42, 50, 54, 56

V.V. Digalakis, D. Rtischev, and L.G. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366, 1995. 52

T.J. Doherty, A.A. Vandervoort, and W.F. Brown. Effects of ageing on the motor unit: a brief review. *Canadian journal of applied physiology*, 18:331–358, 1993. 11

D. Mysak Edward. Pitch duration characteristics of older males. *Journal of Speech and Hearing Research*, 2:46–54, 1959. 12

W. Endres, W. Bambach, and G. Flösser. Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America*, 49 (6B):1842–1848, 1971. 12

P.L. Enright, R.A. Kronmal, T.A. Manolio, M.B. Schenker, and R.E. Hyatt. Respiratory muscle strength in the elderly. Correlates and reference values. *American Journal of Respiratory and Critical Care Medicine*, 149:430–438, 1994. 7

Carole T. Ferrand. Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, 16:480–487, 2002. 15

Eric Fosler-Lussier and Nelson Morgan. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2-4):137–158, 1999. 98, 136

M.J.F Gales. Maximum Likelihood Linear Transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998. Speech recognition systems;. 51, 52

M.J.F. Gales. Cluster Adaptive Training of Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417 – 428, 2000. ISSN 1063-6676. 57

M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10(4):249–264, 1996. 51

G. Garau, S. Renals, and T. Hain. Applying Vocal Tract Length Normalization to Meeting Recordings. In *Proceedings of Interspeech*, pages 265–268, 2005. 47

J.-L. Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994. ISSN 1063-6676. 54

Kallirroi Georgila, Maria Wolters, Vasilis Karaiskos, Melissa Kronenthal, Robert Logie, Neil Mayo, Johanna Moore, and Matt Watson. A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008. 65

Kallirroi Georgila, Maria Wolters, Johanna D. Moore, , and Robert Logie. The MATCH corpus: A corpus of Older and Younger users' interactions with Spoken Dialogue Systems. *Language Resources and Evaluation, to appear*, 2009. 64

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of asr technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, WOCCI '09, pages 7:1–7:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-690-8. doi: http://doi.acm.org/10.1145/1640377.1640384. URL `http://doi.acm.org/10.1145/1640377.1640384`. 2

L. Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP*, volume 1, pages 532–535, May 1989. doi: 10.1109/ICASSP.1989.266481. 82

I.J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 27(10):1032–1043, 1953. 40

T. Hain, P. Woodland, T. Niesler, and E. Whittaker. The 1998 HTK system for transcription of conversational telephone speech. In *IEEE ICASSP*, volume 1, pages 57–60, 1999. 47

T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005a. 64, 68

Thomas Hain, John Dines, Giulia Garau, Martin Karafiat, Darren Moore, Vincent Wan, and Steve Renals. Transcription of conference room meetings: an investigation. In *Interspeech*, 2005b. 124

Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, David van Leeuwen, Mike Lincoln, and Vincent Wan. The 2007 AMI(DA) System for Meeting Transcription. In *Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science*, volume 4625, pages 414–428, 2008. 35, 125

J.D. Harnsberger, R. Shrivastav, W.S. Jr Brown, H. Rothman, and H. Hollien. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice*, 22(1):58–69, 2008. 17

Fredric J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. In *Proceedings of IEEE*, 1978. 25

Trey Hedden and John D. E. Gabrieli. Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews Neuroscience*, 5:87–96, 2004. 11

Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990. 21, 26

James Hillenbrand and Robert A. Houde. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research*, 39:311–321, 1996. 15, 88

James Hillenbrand, Ronald A. Cleveland, and Robert L. Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, 37:769–778, 1994. 15

M Hirano, S Kurita, and S Sakaguchi. Ageing of the vibratory tissue of human vocal folds. *Acta Otalaryngologica*, 107:428–433, 1989. 9

Mei-Yuh Hwang and Xuedong Huang. Subphonetic modeling with Markov states - Senone. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 33–36, 1992. 37

ISO-226:. Acoustics – normal equal loudness level contours, 2003. 21

Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *ICSLP*, 1998. 65

Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyoshiro Shikano, and Shuichi Itahashi. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of the Acoustical Society of Japan*, 20(3):199–206, 1999. 65

F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of IEEE*, 64(4):532–556, 1976. 28

Frederick Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, 1980. 42

Daniel Jurafsky and James H. Martin. *Speech and Language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2 edition, 2008. 41

J.C. Kahane. *Aging Communication processes and disorders*, chapter Anatomic and physiologic changes in the aging peripheral speech mechanism, pages 21–45. Grune & Stratton, Incorporated, 1981. 7

J.C. Kahane and J. Hammons. *Laryngeal Function in Phonation and Respiration.*, chapter Developmental changes in the articular cartridge of the human cricoarytenoid joint, pages 14–28. San Diego, College Hill Press, 1987. 9

George Karypis. *CLUTO : A Clustering Toolkit*, 2003. 108

S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35:400–401, 1987. 40

Tatsuya Kawahara, Tetsunori Kobayashi, Kazuya Takeda, Nobuaki Minematsu, Katsunobu Itou, Mikio Yamamoto, Atsushi Yamada, Takehito Utsuro, and Kiyohiro Shikano. Japanese Dictation Toolkit: Plug-and-Play framework for speech recognition R&D. In *ASRU*, pages 393–396, 1999. 76, 77

K. Kevin and D.R. Philips. Global aging: The challenge of success. *Population Reference Bureau*, Vol 60, No 1, 2005. 1

R.J. Kilch. Relationships of vowel characteristics to listener ratings of breathiness. *Journal of Speech and Hearing*, 25:574–580, 1982. 15

Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184, 1995. 41

Marcel Kockmann, Lukas Burget, and Jan Cernocky. Brno University of Technology system for Interspeech 2010 paralinguistic challenge. In *Interspeech*, 2010. 60

H Koshino, T Hirai, T Ishijima, and Y Ikeda. Tongue motor skills and masticatory performance in adult dentates, elderly dentates, and complete denture wearers. *The journal of prosthetic dentistry*, 77:147–152, 1997. 11

R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695 – 707, 2000. ISSN 1063-6676. 58, 119, 130

Akinobu Lee, Tatsuya Kawahara, and Shuji Doshita. An efficient two-pass search algorithm using word trellis index. In *International Conference on Spoken Language Processing*, pages 1831–1834, 1998. 77

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius an open source real-time large vocabulary recognition engine. In *Eurospeech*, pages 1691–1694, 2001. 45

L. Lee and C. Rose. Speaker normalisation using efficient frequency warping procedures. In *IEEE ICASSP*, pages 353–356, 1996. 47

C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995a. ISSN 0885-2308. 50

C.J. Leggetter and P.C. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. *Proc ARPA Spoken Langauge Technology Workshop*, 9: 171–185, 1995b. 50

N. Levinson. The Wiener RMS error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(4):261–278, 1947. 27

S.E Linville. *Voice Quality Measurement*, chapter The Aging Voices, pages 359–376. Singular Thomson Learning, 2000. 12, 13, 15

S.E Linville. *Vocal Aging*. Singular Thomson Learning, San Diego, 2001. 2, 6, 9, 10, 14, 17

S.E Linville. The aging voice. *The ASHA Leader*, 21:12–14, 2004. 7, 9

Sue Ellen Linville. Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice*, 16:472–479, 2002. doi: 10.1016/S0892-1997(02)00122-4. 17

Julie M. Liss, Gary Weismer, and John C. Rosenbek. Selected acoustic characteristics of speech production in very old males. *Journal of Gerontology*, 45:2, 1989. 95

Patricia Lynne-Davies. Influence of age on the respiratory system. *Geriatrics*, 32: 57–60, 1977. 7

D. Mahler, R. Rosiello, and J. Loke. The aging lung. *Clinics in geriatric medicine*, 2: 215–225, 1986. 7

J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(5): 561580, 1975. 27

H.B. Mann and D.R. Whitney. On a test whether one of two random variables is stochastically larger than the other. In *Ann. Math. Statistics*, volume 18, pages 50–60, 1947. 69

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, and Yoshitaka Hirano. Japanese morphological analysis system ChaSen version 2.0 manual. Technical Report NAIST-ISTR99009, Nara Institute of Science and Technology, 1999. 77

John McDonough, William Bryne, and Xiaoqiang Luo. Speaker normalisation with all pass transforms. In *Proceedings of ICSLP*, 1998. 48

P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. In C. H. Chen, editor, *Pattern recognition and artificial intelligence*, pages 374–388. Academic Press, New York, 1976. 21

Florain Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Fleix Burkhrdt, Joachim Stegmann, Christian Müller, Richard Huber, Bernt Andrassy, Josef Bauer, and Bernard Littel. Comparison of four approaches to age and gender recognition for telephone applications. *ICASSP*, 4:1089–1092, 2007. 60, 107

Nobuaki Minematsu, Mariko Sekiguchi, and Keikichi Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 137–140, Orlando, FL, 2002. 59

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted automata in text and speech processing. In *12th biennial European Conference on Artificial Intelligence*, 1996. 44

Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6): 453–467, 1990. 82

C Müller and F Burkhardt. Combining short-term cepstral and long-term prosodic features for automatic recognition of speaker age. In *Interspeech*, 2007. 59

Christian Müller. Automatic recognition of speakers' age and gender on the basis of empirical studies. *Proc. of Interspeech*, 2006. 59

Christian Müller, Frank Wittig, and Jorg Baus. Exploiting speech for recognising elderly users to respond to their special needs. *Eurospeech*, pages 1305–1308, 2003. 1, 59

Kevin P. Murphy. Hidden semi-Markov models (HSMMs). Technical report, University of British Columbia, 2002. URL `http://www.cs.ubc.ca/~murphyk/Papers/segment.pdf`. 116

M. Nakayama. Histological study on aging changes in the human tongue. *Nippon Jibiinkoka Gakkai kaiho*, 94:541–555, 1991. 11

H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1 – 38, 1994. ISSN 0885-2308. 37, 41

Y Normandin. *Hidden Markov Models, maximum mutual information estimation and the speech recognition problem*. PhD thesis, McGill University, 1991. 28

K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda. Hidden semi-markov model based speech recognition system using weighted finite-state transducer. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages I–I, 2006. 116

Friedrich P. Paulsen and Bernhard N. Tillmann. Degenerative changes in the human cricoarytenoid joint. *Archives of otolaryngology, head & neck surgery*, 124:903–906, 1998. 8

Michael Pitz, Sirko Molau, Ralf Schluter, and Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space. In *Eurospeech*, 2001. 48

D. Povey and P.C Woodland. Minimum phone error and I-smoothing for improved discriminative training. In *IEEE International conference on Acoustics, Speech and Signal Processing*, 2002. 28

M. L. Pretterklieber. Functional anatomy of the Human Intrinsic Laryngeal Muscles. *European Surgery*, 35:250–258, 2003. 8

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series, 1993. 5

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 86, 1989. ISSN 0018-9219. 30, 100

L.A Ramig and R.L Ringel. Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech and hearing Research*, 26:22–30, 1983. 2, 12, 14, 15

Lorraine Olson Ramig, Steven Gray, Kristin Baker, Kim Corbin-Lewis, Eugene Buder, Erich Luschei, Hillary Coon, and Marshall Smith. The aging voice: A review, treatment data and familial and genetic perspectives. *Clinical Linguistics and Phonetics*, 53:252–265, 2001. 2, 7, 12

Steve Renals and Thomas Hain. Speech recognition. In Alex Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Wiley Blackwell, 2010. 2

D.W Robinson and R.S Dadson. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166–181, 1956. 21, 26

M.T. Rodeño, J.M. Sánchez-Fernández, and J.M. Rivera-Pomar. Histochemical and morphometrical ageing changes in human vocal cord muscles. *Acta Otolaryngologologica*, 113:445–449, 1993. 9

A Rossi, A Ganassini, C Tantucci, and V Grassi. Aging and the respiratory system. *Aging*, 8:143–161, 1996. 7

Paul Rother, Balthasar Wohlgemuth, Werner Wolff, and Ines Rebentrost. Morphometrically observable aging changes in the human tongue. *Annals of Anatomy*, 184: 159–164, 2002. 11

K. Sato and M. Hirano. Age-related changes of elastic fibers in the superficial layer of the lamina propria of vocal folds. *The Annals of otology, rhinology, and laryngology*, 106:44–48, 1997. 9

T. Sato and H. Tauchi. Age changes in human vocal muscle. *Mechanisms of Ageing and Development*, 18:67–74, 1982. 9

Susanne Schötz. A perceptual study of speaker age. Technical report, Lund University, Dept. of Linguistics, 2001. 107

Susanne Schötz and Christian Müller. *Speaker Classification II*, chapter A Study of Acoustic Correlates of Speaker Age, pages 1–9. Springer Berlin / Heidelberg, 2007. 12, 17

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. The Interspeech 2010 paralinguistic challenge. In *Interspeech*, 2010. 60

I. Shafran, M. Riley, and M. Mohri. Voice signatures. In *IEEE Automatic Speech Recognition and Understanding workshop*, pages 31–36, St Thomas, VI, USA, 2003. 58, 59

C.E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948. 42

K. Shinoda and C.-H. Lee. A structural Bayes approach to speaker adaptation. *Speech and Audio Processing, IEEE Transactions on*, 9(3):276 –287, mar. 2001. ISSN 1063-6676. doi: 10.1109/89.906001. 55

K. Shinoda and Chin-Hui Lee. Structural MAP speaker adaptation using hierarchical priors. In *Automatic Speech Recognition and Understanding Proceedings*, pages 381–388, 1997. 55

Elaine M. Shuey. Intelligibility of older versus younger adults' cvc productions. *Journal of Communication Disorders*, 22:437–444, 1989. 60

O. Siohan, C. Chesta, and Chin-Hui Lee. Joint Maximum A Posteriori adaptation of transformation and HMM parameters. *IEEE Transactions on Speech and Audio Processing*, 9(4):417–428, may. 2001. ISSN 1063-6676. 56

Olivier Siohan, Tor Andre Myrvoll, and Chin-Hui Lee. Structural Maximum A Posteriori Linear Regression for fast HMM adaptation. *Computer Speech and Language*, 16(1):5 – 24, 2002. ISSN 0885-2308. 56

Stanley Smith Stevens. On the psychophysical law. *Psychological review*, 64:153–181, 1957. 27

Stanley Smith Stevens, John Volkman, and Edwin Newman. A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937. 21, 25

Andreas Stolcke, Luciana Ferrer, Sachin Kajarekar, Elizabeth Shriberg, and Anand Venkataraman. MLLR transforms as features in speaker recognition. In *Interspeech*, pages 2425–2428, 2005. 52, 107, 120

Andreas Stolcke, Luciana Ferrer, and Sachin Kajarekar. Improvements in MLLR-transform-based speaker recognition. In *Odyssey*, pages 1–6, 2006. 120

Paul Taylor, Alan W. Black, and Richard Caley. The architecture of the Festival speech synthesis system. In *The Third ESCA Workshop in Speech Synthesis*, pages 147–151, 1998. 68

K. Tolep, N. Higgins, S. Muza, G. Criner, and S.G. Kelsen. Comparison of diaphragm strength between healthy adult elderly and young men. *American Journal of Respiratory and Critical Care Medicine*, 152:677–682, 1995. 7

L.F. Uebel and P.C. Woodland. An investigation into vocal tract length normalisation. In *Sixth European Conference on Speech Communication and Technology*, 1999. 48

Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260–269, 1967. 32

Barbara Weinstein. *Geriatric Audiology*, chapter The Biology of Aging, pages 15–40. Georg Thieme Verlag, 2000. 10

Angie Williams and Howard Giles. Sociopsychological perspectives on older people's language and communication. *Ageing and Society*, 11:103–126, 1992. 60

Jay G. Wilpon and Claus N. Jacobsen. A study of speech recognition for children and the elderly. In *ICASSP*, volume 1, pages 349 – 352, Atlanta, GA, USA, 1996. 2, 61, 134

Maria Wolters, Kalliroi Georgila, Robert Logie, Sarah MacPherson, Johanna D. Moore, and Matt Watson. Reducing working memory load in spoken dialogues. *Interacting with Computers*, 21:276–287, 2009. 2, 64

Jian Wu and Eric Chang. Cohorts based custom models for rapid speaker and dialect adaptation. In *Eurospeech*, pages 1261–1264, 2001. 120

S.A. Xue and G.J. Hao. Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *Journal of Speech, Language, and Hearing Research*, 46(3):689 – 701, 2003. ISSN 1092-4388. 10

Steve An Xue and Dimitar Deliyski. Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology*, 27:159–168, 2001. 12, 13, 15

S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, A. Lee, and K. Shikano. Unsupervised training of phoneme models using HMM sufficient statistics and a speaker distance function. *Electronics and Communications in Japan, Part 3 (Fundamental Electronic Science)*, 88(9):33 – 41, 2005. ISSN 1042-0967. 120

S. Young, N. Russell, and J. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical report, Cambridge University Engineering Department, 1989. 45

S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Workshop on Human Language Technology*, pages 307–312, 1994. 37

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for Hidden Markov Model Toolkit Version 3.4)*, 2006. 21, 36, 67, 77