Speaker Characterization using Adult and Children's Speech

by

Saeid Safavi

A thesis submitted to
The University of Birmingham
for the degree of

Doctor of Philosophy

School of Electronic, Electrical and Systems Engineering
The University of Birmingham
June 2015

Thesis advisors: Professor Martin Russell & Dr. Peter Jancovic Saeid Safavi

Speaker Characterization using Adult and Children's Speech

Abstract

Speech signals contain important information about a speaker, such as age, gender, language, accent, and emotional/psychological state. Automatic recognition of these types of characteristics has a wide range of commercial, medical and forensic applications such as interactive voice response systems, service customization, natural human-machine interaction, recognizing the type of pathology of speakers, and directing the forensic investigation process. Many such applications depend on reliable systems using short speech segments without regard to the spoken text (text-independent). All these applications are also applicable using children's speech.

This research aims to develop accurate methods and tools to identify different characteristics of the speakers. Our experiments cover speaker recognition, gender recognition, age-group classification, and accent identification. However, similar approaches and techniques can be applied to identify other characteristics such as emotional/psychological state. The main focus of this research is on detecting these characteristics from children's speech, which is previously reported as a more challenging subject compared to adult. Furthermore, the impact of different frequency bands on the performances of several recognition systems is studied, and the performance obtained using children's speech is compared with the corresponding results from experiments using adults' speech.

Speaker characterization is performed by fitting a probability density function to acoustic features extracted from the speech signals. Since the distribution of acoustic features is complex, Gaussian mixture models (GMM) are applied. Due to lack of data, parametric model adaptation methods have been applied to adapt the universal background model (UBM) to the characteristics of utterances. An effective approach involves adapting the UBM to speech signals using the Maximum-A-Posteriori (MAP) scheme. Then, the Gaussian means of the adapted GMM are concatenated to form a Gaussian mean super-vector for a given utterance. Finally, a classification or regression algorithm is used to identify the speaker characteristics. While effective, Gaussian mean super-vectors are of a high dimensionality resulting in high computational cost and difficulty in obtaining a robust model in the context of limited data. In the field of speaker recognition, recent advances using the i-vector framework have increased the classification accuracy. This framework, which provides a compact representation of an utterance in the form of a low dimensional feature vector, applies a simple factor analysis on GMM means. Motivated by this success, the i-vector framework is applied to the age-group, gender and accent recognition problems and the performances are compared with the corresponding results form different acoustic frameworks. In these approaches, each utterance is modeled by its corresponding i-vector. Then Linear Discriminant Analysis (LDA) is used for minimizing within class variability while maximizing between class variability. Finally, a Support Vector Machine (SVM) classifier or simple dot product is applied to estimate the characteristic of speakers.

A new analysis for investigating the importance of different parts of the speech spectrum for speaker, accent, gender, and age-group identification are proposed. For speaker identification, the practically important problems of identifying a child in a simulated class and school of children are studied. The results of gender identification using children's speech demonstrate the effects of age and puberty on the performance of gender identification systems. For having a baseline to compare the performance of automatic speaker characterization systems, when using children's speech, human experiments are conducted for gender and age-group identification tasks.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms	**Definition**
ABI	The Accents of the British Isles
AG	Age Group
Age-ID	Age Identification
AID	Accent IDentification
ANN	Artificial Neural Networks
ASV	Automatic Speaker Verification
BCU	Birmingham City University
BL	Band Limited
CMS	Cepstral Mean Subtraction
DCT	The Discrete Cosine Transform
DET	Detection Error Tradeoff
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
FB	Full Bandwidth
FBE	Filter Bank Energies
FFT	Fast Fourier Transform
FW	Feature Warping
GI	Gender Identification
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICASSP	International Conference on Acoustics, Speech and Signal Processing

ICSLP	International Conference on Spoken Language Processing
IEEE	Institute of Electrical and Electronics Engineers
ISV	Intersession Variability
JFA	Joint Factor Analysis
LBG	Linde-Buzo-Gray algorithm
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coefficient
MAP	Maximum A-posteriori Probability
MFCC	Mel Frequency Cepstral Coefficient
MIT	Massachusetts Institute of Technology
ML	Maximum Likelihood
MMI	Maximum Mutual Information
MVN	Mean and Variance Normalization
NAID	Normalized Accent Identification
NIST-SRE	National Institute of Standards and Technology-Speaker Recognition Evaluation
NSID	Normalized Speaker Identification
PCA	Principal Component Analysis
PDF	Probability Density Function
PRLM	Phone Recognition and Language Modelling
PPRLM	Parallel Phone Recognition and Language Modelling
QP	Quadratic Programming
RS	Recognition Systems
SAD	Speech Activity Detector
SDC	Shifted Delta Cepstral
SID	Speaker Identification
SPA	Sailor Passage A
S-space	Super-vector space
SRE	Speaker Recognition Evaluation
SR	Speaker Recognition
SSE	Standard Southern English
SVM	Support Vector Machine
SVR	Support Vector Regression

TI	Texas Instrument
T-Matrix	Total Variability Matrix
T-Norm	Test Normalization
TS	Test Set
TRAPs	Temporal Patterns
UBM	Universal Background Model
VQ	Vector Quantization
WOCCI	Workshop on Child Computer Interaction
WSNMF	Weighted Supervised Non-Negative Matrix Factorization
Z-Norm	Zero Normalization

FOR MY DEAR PARENTS WHO TAUGHT ME HOW TO LOVE, UNCONDITIONALLY, TO FORGIVE AND TO FORGET, ALWAYS

Acknowledgments

The completion of this research would never have been possible without the assistance and support of the many faculty, family, and friends who have been a part of my life over the past years.

I owe an infinite debt of gratitude to my supervisor, Prof. Martin Russell, who was abundantly helpful and offered invaluable assistance, support, and guidance. I specially wish to thank him for taking the time and effort to read through multiple drafts and provide suggestions to improve my research and my writing; without whose knowledge and assistance this study would not have been successful. Sincere gratitude is also due to my advisers, Prof. Mike Carey and Dr. Peter Jancovic, for their unfailing support throughout my research studies.

I also wish to extend my gratitude to Dr. Soud Hanani, who constantly shared with me his precious definition of life and many practical tips to enhance my research experience.

I extend my most heartfelt thanks to my parents and sister. The sacrifices they have made for my benefit can never be repaid. I thank them most for their love, understanding, and patience and I hope I can someday adequately express my appreciation.

At the end, I would also like to appreciate my friends; Mandana, Maryam, Dina, Donya, Niloufar, Emilie, Eva, Linxue, Chloe, Sara, Morvarid, Nikolina, Mersedeh, Phil, Mehdi, Mohammad Ali, Amir, Kasra, Sam, Ashkan, Roozbeh, Alborz, Masoud, Steve, Mahran, and Mohammad, for supporting me in every possible way.

"Now it is established in the sciences that no knowledge is acquired save through the study of its causes and beginnings, if it has had causes and beginnings; nor completed except by knowledge of its accidents and accompanying essentials."

Avicenna (Ibn Sīnā)

1

Introduction

1.1 Speech Technologies for Children

Some of the most compelling applications of spoken language technology in education involve children, but computer recognition of children's speech is particularly difficult. This is due to the fact that children have large differences in both the acoustic and the linguistic aspects of speech compared to adults [9–11]. Differences in the pitch, the formant frequencies, the average phone duration, the speaking rate, the glottal flow parameters, pronunciation and grammar are the various acoustic and linguistic differences [12, 13]. As reported, children have different values of mean and variance of the acoustic features of speech than those of adults [10, 14, 15]. For example, the area of the F1-F2 formant ellipses is larger for children than for adults for most vowel phonemes and children speech contains more dis-fluencies and extraneous speech [16].

As all children undergo rapid development and with varying rates, it is difficult to model their constantly changing speech characteristics. Also as children grow, their speech production organs change and so their anatomy and physiology keep changing quite significantly. Thus, comparing with those of adults speech, children's speech has higher inter- and intra-speaker acoustic variabilities [11, 17].

The acoustic and linguistic characteristics are two types of properties of a speech signal. The physiology of the speaker governs the acoustic properties of a speech signal. The physical dimensions of vocal tracts and vocal folds have direct influence on the acoustics of speech signal. What makes children's speech different from adults' speech has been examined by several researchers. Age-dependent changes in the formant frequencies and the fundamental frequency measurements of children speakers aged three to thirteen were reported by Eguchi and Hirsh [18], and later summarized by Kent [19]. Compare to adults, children have formants located at higher parts of the spectrum [10, 12, 17, 20]. Also, they have high pitch frequency values which cause large spacing between the pitch harmonics [10, 12, 17, 21]. These high formant frequencies and pitch frequency values are attributed to their inherent shorter vocal tract and vocal folds lengths, respectively. Researchers have also observed that the phoneme durations and the average sentence durations are longer than those for adults, which consequently reduces their speaking rate [10, 12, 17, 22]. On studying consonant-vowel transitions in the case of adults' and children's speech [23], it is reported that the children's speech has shorter transition duration and larger spectral difference between consonant and vowel in the consonant-vowel pair than those of adults' speech.

On the other hand, due to physiological differences between the speakers, the differences in the voice source parameters affect the source spectrum [24]. Contributing to all this is their increased intra-speaker spectral and temporal variabilities [9, 17, 23]. Increases in the intra-speaker spectral and temporal variabilities cause greater overlapping of the phonemic classes making the pattern classification problem even more difficult.

The usual legal definition for child is anyone under 18, but in this study the focus is on the 5 years to 16 years old children. In general 5-year old children are able to say all speech sounds in words, although they may make mistake on sounds that are harder to say. At this stage they can talk (including names, letters and numbers) without repeating sounds or words most of the time [25].

In a report, it has been stated that the automatic systems when using speech signals from 5-year old children have about 60% of vowel classification accuracy against that of about 90% when using the adult speech signals [10]. Two vowel classification methods are studied and both methods involve the computation of a variance-normalized distance measure between the feature vector for a given token and the centre (or centroid) of each of the training categories (they used 10 vowel categories). Overall classification accuracy is determined by calculating the percentage of tokens that were assigned by the classifier to the category that was intended by the talker.

Mentioned earlier in this chapter, in addition to acoustic differences, children have large differences in linguistic correlates of speech. Children display less precise control of the articulators especially at the age of 5-6 years (the special interest on 5 and 6 years old children is due to the fact that they usually are the youngest participants in speech related experiments). Consequently, children's speech have many problems such as dis-fluencies, false-starts and extraneous speech [9, 26, 27]. As children have smaller vocabulary than adults, they use fewer words per utterance to convey the same message. And occasionally their sentences have some spurious words which are not found in adults' case. To convey the intended message, older children use simpler linguistic constructs. So, the ability of children to use language efficiently to convey the message depends on their age.

1.2 Speaker Characterization Technologies for Children

While speech recognition is dealing with extracting the underlying linguistic message in an utterance, speaker characterization is preoccupied with extracting characteristics of the speaker who is speaking the utterance. In addition to its linguistic content, speech signals also carry important paralinguistic information about a speaker such as identity, age, gender, language, accent and emotional/psychological state. Although automatic recognition of children's speech, and variability of acoustic parameters of speech as a function of age has been the subject of considerable research effort, there is little published work on issues and algorithms related to automatic recognition of a child's paralinguistic characteristics from his or her speech. For example, how increases in inter- and intra-speaker variability for children's speech [28] will affect the performance of the automatic speaker characterization system is unknown.

As it is mentioned earlier, acoustic and linguistic characteristics of children's speech are different from those for adult, therefore for children's speech, the influence of bandwidth reduction on speech recognition accuracy is greater than for adults [29, 30]. Although the relevant studies for adult speaker recognition have been reported [31], the significance of different frequency bands for automatic recognition of children's characteristics is unknown.

In several application areas, including, security, child protection, and education, the employment of speaker characterization technology for children could be helpful. For instance it could be useful for teenagers and young adults aged from 8 to 17 who are using internet and wanted to set up their own profile on a social networking site. A valuable safeguard for a child engaged in social networking are speaker, age and gender recognition systems that identifies a child's identity, gender and age-group based on his or her voice, and confirms the identity of the individual with whom the child is communicating. Another example is an interactive educational tutor which will able the child to login by his/her voice, with no need to go through a formal login process, the tutor could automatically identify a child in a class to continue a previous lesson, adjust its content to suit the child's age-group and gender, and process the child's responses to provide relevant feedback.

The main focus of this research is to study automatic speaker characterization technologies for child speech. To obtain a base line system, we initially designed a system to address the problem of speaker and accent recognition, in both clean and noisy environments, for adults' speech. Then the modified speaker recognition system (based on adult speech) is used to study automatic characterization of children's speech. Among all speaker characteristics, we study identification of speaker's identity, gender, and age-group, from children's speech (and in some cases also with adult speech to be able to compare the performance).

1.3 Review of Speech Detection Technologies for Adults

The general area of speaker recognition is divided into verification and identification. The focus of this research are both verification and identification, in which the goal is to determine from a voice sample if a person is who he or she claims (verification) and is to determine an unknown speaker's identity (identification).

Automatic recognition of speaker characteristics has a wide range of commercial, medical and forensic applications in real-world scenarios. For example, in a multilingual call-centre [32], a call should be directed to an agent whose language matches the customer. To find the best agent for a call, an automatic dialect/accent recognition system can be considered to avoid typical misunderstandings in the agent-customer conversation [33]. In this case, automatic age estimation can also be applied as elderly customers usually prefer an agent with a slow speech rate [34]. Targeted advertising through the Internet, where user-computer and user-company vocal interaction has increased significantly during the last decades, is another scenario of application. In this case, information about the user's language/accent, age and gender can help to offer appropriate products and services. In video games, knowledge about a user's characteristics can help the game to adapt itself. For example, the preference for the game music might differ significantly between a male teenager compared to an adult female. Speaker characterization is also applied to diagnosis, analysis and monitoring of different diseases such as autism and Parkinson's disease.

In addition, automatic identification of speaker characteristics can improve the performance of automatic speech recognition (ASR) systems. A fundamental challenge of using ASR systems in real world markets such as telephone networks and personal computers is their significant performance drop for some speakers. As speech interaction with computers becomes more pervasive, and its applications (such as telephone financial transactions and information retrieval from speech databases) become more private and sensitive, there is a growth in the value of automatic recognition of a speaker based on vocal characteristics.

This research aims to develop accurate methods and tools to identify different characteristics of the speakers, using both adult and child speech. Automatic recognition systems should be robust enough for facing possible environmental variability, such as transmission over a communication channel, and background noise. Moreover, they must be capable of achieving high accuracy for short input speech samples. From the practical point of view there is usually no control over the duration of speech. It is important to be able to identify a speaker using a short speech segment with acceptable accuracy level. From a theoretical point of view, it is interesting for text-independent operation to determine if the system can characterize a speaker's voice well enough using the short segment of speech.

1.4 Importance of Different Parts of the Spectrum for Speech Technologies

Currently the most commonly used parameterization is to represent a spoken utterance as a sequence of Mel Frequency Cepstral Coefficient (MFCC) vectors, which are calculated using the entire frequency bandwidth. However, we know that different frequency regions contain different types of information. For instance a study for speaker identification [31], showed that the frequency regions below 600 Hz and above 3000 Hz provided better speaker identification accuracy than the middle-frequency regions, when using mono Gaussian modelling and the TIMIT corpus. However, no similar study has been reported for other types of automatic speaker characterization using adult speech.

Russell et al. [35] show that the bandwidth affects recognition of children's speech by humans and machines. The effect of bandwidth reduction from 8 to 4 kHz was reported, which shows that this effect is over 100% greater for children's speech than for adults. However, no similar studies have been reported on the effect of different frequency sub-bands on the performance of automatic speaker characterization systems using children's speech.

1.5 Scope of the Thesis

Figure 1.5.1 shows the overall overview of this thesis. The thesis is organized as follows:
Chapter 2 provides a background review of techniques for the area of automatic identification of speaker characteristics. The chapter begins by reviewing some of the early experiments which examined how humans recognize a speaker's identity. This is followed by an outline for some of the major feature sets used for identification of speaker characteristics, such as speaker's identity, age and gender. Finally, an outline of some of the major techniques, for modelling, normalization and scoring, used for identification of speaker characteristics are presented and notation and nomenclature used throughout the thesis are introduced.
Chapter 3 contains the description of corpora, for both adult and child speech, which are used for experiments of following chapters.
Chapter 4 presents the relevant experiments for obtaining the baseline systems for both telephony and microphone recorded speech signals. These experiments are related to the effect of using different feature sets and classification techniques. Based on the result of experiments presented in this chapter automatic speaker characterization systems are designed and presented on each of the following chapters.
Chapter 5 introduces the contrastive effects of different frequency bands on the performance of speaker and and accent identification, using clean recorded speech from adult speakers. In addition for both speaker and accent identification, the best obtained performances are presented in this chapter.
Chapter 6 studies the speaker recognition tasks for children's speech. However in order to have a baseline to compare the performance of automatic recognition system using children's speech, some experiments with the comparable experimental configuration are also conducted using adult speech recordings. In addition, the regions of the spectrum carrying more speaker relevant information are identified and studied, for both child and adult speech, using two methods.
Chapter 7 discuses several approaches to automatic identification of gender and age-group

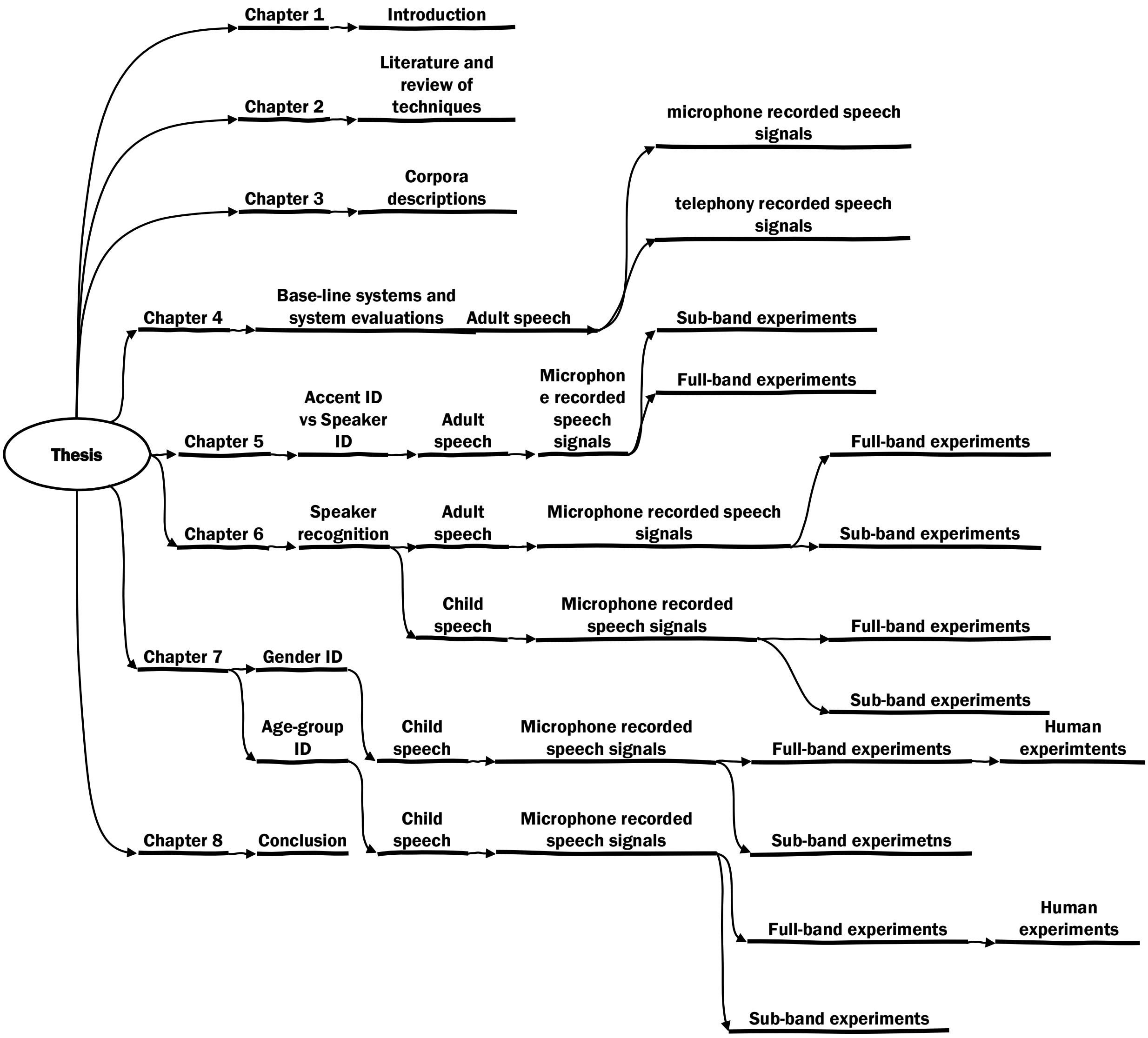


Figure 1.5.1: Overview of whole thesis.

from children's speech. The effect of different frequency bands on the performance of automatic gender and age-group identification systems are studied in this chapter. In addition, the identification rate from human experiments are presented for both gender and age-group identification tasks.

Finally, **Chapter 8** summarizes the major results and conclusions of the thesis and suggests future directions for research.

1.6 Major Contributions

The research described in this thesis provides original contributions to the field of automatic recognition of speaker's identity, gender, and age-group, using children's speech. The major contributions can be summarised as follows:

1. A new analysis of the importance of different parts of the speech spectrum for speaker and accent identification, using adult's speech. (Published as an IEEE signal processing letter in 2012, [36]. Presented at ICASSP 2012 and UK Speech 2012.)

2. The first evaluation of the utility of current speaker recognition techniques for children's speech. These experiments show how speaker recognition performance depends on age and compares the regions of the spectrum that are most important for speaker recognition for children and adults. (Published in Interspeech 2012 and WOCCI 2014, [37, 38]. Presented at Interspeech 2012 and WOCCI 2014.)

3. The application of speaker identification to the practically important problems of identifying a child in a simulated class and school of children. Demonstration of the effects of age on performance. (Published in Interspeech 2012, [37]. Presented at Interspeech 2012 and Birmingham City University (BCU) in 2013.)

4. The first evaluation of identification of gender from children's speech, by humans and machines. The results demonstrate the effects of age and puberty on the performance of gender identification systems. In addition, the utilities of different parts of the spectrum are studied. (Published in Interspeech 2013, [39]. Presented at Interspeech 2013 and UK Speech 2013.)

5. Initial evaluation of age-group identification, by humans and machines, from children's speech. The effect of gender on the performance of age-group identification task is studied, and useful bands for this task are identified. (Published in Interspeech 2014, [40]. Presented at Interspeech 2014 and UK Speech 2014.)

"Absence of understanding does not warrant absence of existence."

Avicenna (Ibn Sīnā)

2

Background and Review of Techniques

This chapter provides a background review of techniques for the area of automatic identification of speaker characteristics. It begins by reviewing some of the early experiments which examined how humans recognize a speaker's identity. This is followed by outline of some of the major techniques, for modelling, normalization and scoring, used for identification of speaker characteristics.

2.1 Speech Production

Figure 2.1.1 shows a general model for human speech production. The production of spoken language involves three major levels of processing: conceptualisation, formulation, and articulation. In the first process, conceptualisation or conceptual preparation, the intent to speak gives rise to an association between a desired notion and a specific word to be conveyed. The second process of formulation involves the creation of the linguistic form needed to convey the chosen message. This process is supported by three mechanisms, namely, grammatical, morpho-phonological, and phonetic encoding. The mechanism of grammatical encoding entails the selection of the most suitable syntactic word or lemma. In the third process, articulation, speech is produced through the specific and interlinked functions of the vocal apparatus components, namely, lungs, glottis, larynx, tongue, lips, and jaw [41].

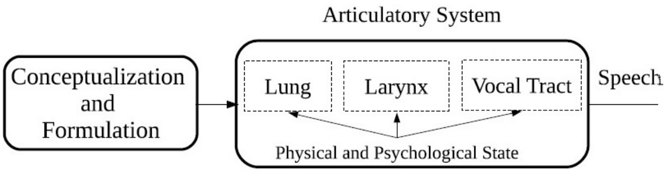

Articulatory System
Conceptualization and Formulation
Lung
Larynx
Vocal Tract
Speech
Physical and Psychological State

Figure 2.1.1: A simplified human speech production model.

Speech production involves development and control of skilled movement patterns. Many studies have investigated the differences between speech production of adults and children [42–44]. Some studies have confirmed that control of the motor system develops progressively in childhood and deteriorates with ageing. Researchers found that when children (aged from 8 to 17) and older adults (aged from 45 to 84) executed limb or speech movement patterns, they tended to demonstrate reduced velocity of movement and greater variability, and their performance was less accurate than that of young adults (aged from 17 to 45) [44–47].

In a study conducted by Sharkey et al. [47], the development of speech motor skills was investigated in adults and in children aged 4, 7, and 10 years. It was observed that variation between children and adults diminished, with respect to age, in terms of the length of actions such as lip-opening and jaw opening movements and lip-open and jaw-open postures, as well as in terms of the duration between the start of lower lip opening and jaw opening. However, within a group of children, these actions did not vary considerably. Furthermore, in comparison to children of other ages, there was a substantial decline in lower lip displacement variation between children aged 4 and aged 7.

Later, in [44], control of the lower lip, jaw, and larynx (i.e., fundamental frequency) was studied using a non-speech visuomotor tracking (VMT) task. In this research, accuracy and within- and between-subject variability in tracking performance were measured. The results confirmed that the performance of the younger adults was better than that of the children and older adults. Accuracy of movement amplitude tended to increase during development and decline with ageing, whereas age did not appear to influence the accuracy of temporal parameters in lip and jaw tracking. In contrast, age tended to influence individual variability in temporal but not amplitude parameters [44].

In a research by MacDonald et al. [43], human-specific vocalizations were characterized as having two main categories: those that appeared during the maturation stage (independent of experience), and those that depended on early daily interaction. Human language falls into this latter class of vocal learning [43].

2.2 Front-end Analysis of Speech Signals

In order to detect a speaker characteristic, understanding the perceptual mechanisms used by human listeners could be useful in designing automatic speaker characterization systems. There are two approaches to feature extraction, knowledge driven and data driven. Most of the knowledge driven based approaches are only based on the structure of human ear and there still is no complete understanding about human auditory system, e.g. brain functions during communication. So for better understanding the combination of knowledge and data driven approaches is useful.

2.2.1 Perceptual cues used for speaker characterization

There have been several approaches aimed at identifying the perceptual cues used by listeners for associating an utterance with a characteristic of a speaker [48–52].

Humans use several levels of perceptual cues for paralinguistic processing of speech signals. Shown in the Table 2.2.1, the hierarchy of perceptual cues was defined by Reynolds and Heck [8] for recognition of speaker's identity. In this research, they have tried to show that although high-level information is hard to extract from the speech samples, it will lead to more robust automatic speaker recognition system, if it could be extracted.

2.2.2 Feature attributes

In research by Wolf [53], the attributes of ideal features for speaker recognition are discussed, these attributes are also desirable for other speaker characterization tasks. Ideally, the selected features should occur naturally and frequently in speech, be easily measurable, not change over time or be affected by the speaker's health, not be affected by reasonable background noise, not depend on specific transmission characteristics, and not be susceptible to mimicry [53].

Table 2.2.1: Hierarchy of perceptual cues [8].* The anatomical structure of vocal apparatus is usually easy to extract automatically, but there are some exceptions; for example it is true of general vocal tract structure but automatic estimation of tongue position during speech is not an easy task.

	Perceptual	**cues**	
High-level (learned traits)	Semantics, diction, pronunciations, idiosyncrasies	Socio-economic status, education, place of birth	**Difficult to automatically extract**
↓ ↓	Prosodics, speed intonation, rhythm, volume modulation	Personality type, parental influence	↓ ↓
Low-level (physical traits)	Acoustic aspect of speech, nasal, deep, breathy, rough	Anatomical structure of vocal apparatus *	**Easy to automatically extract**

In practice, it is highly improbable to find any set of features which concurrently have all these attributes. Depending on the application, partial or total relaxation of some of these standards will be necessary.

2.2.3 Features for automatic speaker characterization

Based on the results of the studies of acoustic correlates to perceptual cues, the primary focus in the search for features for automatic speaker characterization systems has been on acoustic parameters, such as measures of the spectrum, and some prosodic features such as pitch contours, formant trajectories and speech event timings (e.g., voice onset). In the linear acoustic model of speech production, the composite speech spectrum consists of excitation signal filtered by a time-varying linear filter representing the vocal tract shape. Linear predictive coefficients (LPC) and their various transformations [54], and filter-bank energies and their cepstral representation [55] are the common spectrum representations. The disadvantage of the LPC representation is that it is based on an all pole filter model which may ignore significant speech spectral characteristics for speech corrupted with noise [56]. The filter-bank energies are direct measurements of the energy in different frequency bands and are not dependent on any model constraints [55]. The Mel-scale filter-bank energies and the related cepstral representation owns most of the characteristics of ideal features for automatic speaker characterization.

Cepstral analysis technique took the lead for feature extraction in the fields of ASR and consequently for automatic speaker characterization. This is mainly because this technique provides

Table 2.2.2: High-level feature in SR. EER stands for Equal Error Rate.

Feature Type	Feature description	Selected reference(s)
Prosodic features	Pitch and energy distributions	[59] Using log-pitch+log energy+ their derivative, achieved EER of 16.3%
	Pitch and energy track dynamics	[59] Using slop+duration achieved EER of 14.1%
	Prosodic statistics	[60] Using 19 statistics, duration & pitched related features, achieved EER of 8.1%
Phone features	Phone *N*-grams	[61] Using 5PPRLM phone streams achieved EER of 4.8%
	Phone binary trees	[62] Using 3 token history (4-grams), achieved EER of 3.3%
	Cross-stream phone modeling	[63] By fusing cross-stream & temporal system achieved EER of 3.6%
	Pronunciation modeling	[64] By comparing word-level phone streams with open-loop streams, achieved EER of 2.3%
Lexical features	Word *N*-grams	[65, 66] Using n-gram idiolect system achieved EER of 9.0%
Conversational features	Turn taking pattern & conversational style	[60] Using conditional word usage results in EER of 26%

a methodology for separating the excitation from the vocal tract shape [57]. The outcome of a cosine transform of the real logarithm associated with the short-term spectrum of energy measured on a Mel-frequency scale is a series of Mel-Frequency Cepstrum Coefficients (MFCC) which offer a compact representation. There are several factors that may adversely influence the efficiency of the MFCC, including filter number, filter shape, the distribution of filters, as well as the distortion of the power spectrum. The 0^{th} coefficient is not included in the standard MFCC calculation. According to Zheng et al., this calculation is useful as it can be considered the equivalent of the generalised Frequency Band Energy (FBE), giving rise to the FBE-MFCC [52, 58].

For speaker characterization, measurements (from spectrum) of nasals and vowels were found to be particularly good [67, 68], this is confirmed by investigating frequency regions which contains more speaker specific information by [31]. Based on the theory of speech production [69], voiced speech is the output of a stream of air puffs from the glottis producing resonances in the vocal tract and nasal cavities. Although the pitch reflects the excitation from the glottis, it is strongly affected by factors other than anatomical structure, like emotional state and speech effort. However the speech spectrum reflects the anatomical structure of a person's vocal tract and nasal cavities and therefore has information about distinctive physical attributes. That is why vowels, which are produced by a fairly fixed vocal tract shape, and nasals mainly produced by resonances of the nasal cavities, were shown to be effective for speaker discrimination [31].

The more recent investigations prove that inserting higher-level cues as features, by either combining or fusing, has the ability to improve accurateness and add robustness. Table 2.2.2 shows the effectiveness of high-level features in the speaker recognition area.

2.2.4 MFCC extraction

As it is shown in Figure 2.2.1, the extraction of MFCC features includes the following steps.

1. A pre-emphasis filter, which is a high pass filter, is applied on the signal $x[n]$ in time domain by using: $y[n] = x[n] - a * x[n-1]$, where a is set to 0.95, to remove the part of the samples that did not change in relation to its adjacent samples,

2. To produce a short time speech segment for analysis, the speech signal is periodically multiplied by a Hamming window with a fixed length,

3. The discrete Fourier spectrum is obtained through a fast Fourier transform (FFT) from which the magnitude squared spectrum is calculated,

4. The result is put through a bank of triangular filters [55]. Critical band filtering with a set of triangular band pass filters, which operate directly on the magnitude spectrum is simulated by the filter-bank,

5. Take the logarithm of all filter-bank energies,

6. The last step is to take the discrete cosine transform (DCT) of the log filter-bank energies to decorrelate the filter-bank energies.

The critical band warping is done following an approximation to the mel-frequency scale which is linear up to 1 kHz and logarithmic above 1 kHz, as it is shown in Figure 2.2.2. The centre frequencies of the triangular filter follow a uniform 100 Mel-scale spacing and the bandwidths are set so the lower and upper passband frequencies of a filter lie on the centre frequencies of the adjacent filters, giving equal bandwidths on the Mel-scale but increasing bandwidths on the linear frequency scale. The number of filters, N_F, is selected to cover the signal bandwidth$[0, fs/2]$ Hz, where *fs* is the sampling frequency.

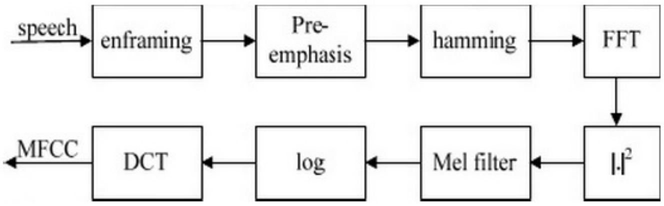


Figure 2.2.1: Feature extraction system.

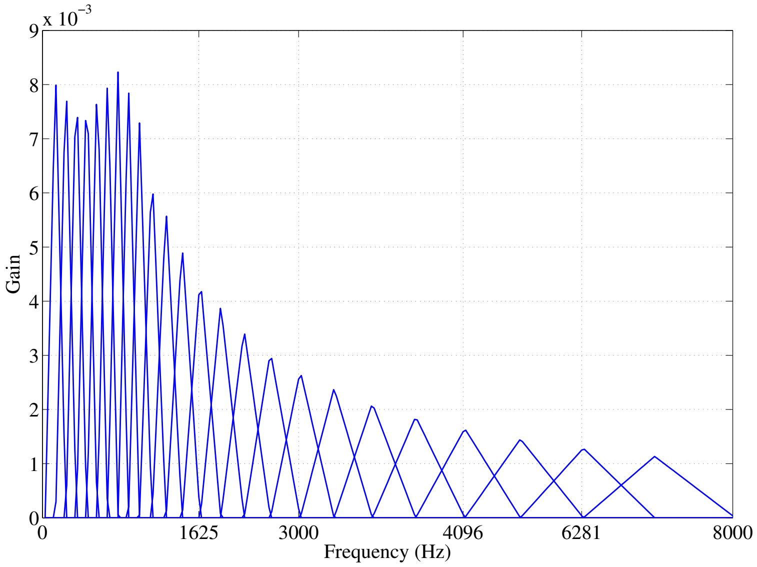


Figure 2.2.2: Mel-scale triangle filter-bank used for feature extraction. In this example the *fs* was equal to 16kHz.

2.2.4.1 Delta coefficient

The performance of a automatic speaker characterization system can be greatly enhanced by adding time derivatives to the basic static parameters.

The delta coefficients are computed using the following regression formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.1}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\Theta}$ to $c_{t+\Theta}$.

2.2.4.2 Shifted-delta cepstral

Shifted-delta cepstral (SDC) coefficients are obtained by linking the delta cepstra computed across multiple frames of speech, as shown in Figure 2.2.3.

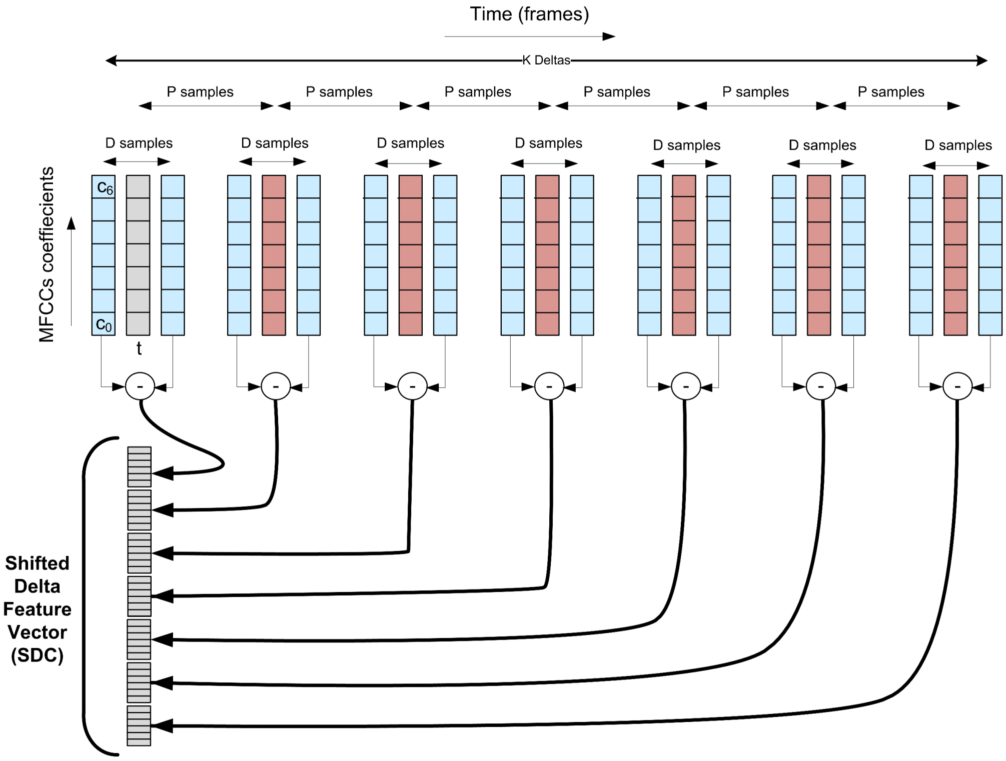


Figure 2.2.3: Block diagram of the process of computing the SDC coefficients [1].

the SDC coefficients for a cepstral frame i at time t, are computed as follow:

$$SDCc_n(t,i) = c_n(t+iP+d) - c_n(t+iP-d) \tag{2.2}$$

where $n = 0, ..., N-1$, and $i = 0, ..., k-1$. n is the n^{th} cepstral coefficient and i is the block number. N is the number of cepstral coefficients computed at each frame, d represents the time advance and delay for the delta computation, k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks [70].

2.2.5 Feature normalization techniques

The noise and channel effects have great impact on most of the speech related technologies. In telephony speaker recognition applications, one of the main decreasing factors for recognition performance, is the channel variability. To standardize distribution parameters of cepstral coefficients over a specified time interval, the feature warping and Mean and Variance Normalization (MVN) [2] algorithms were proposed. Cepstral Mean Subtraction (CMS) [71] and Rasta filtration [72] are other common techniques which could be mentioned. The subsequent two subsections give a brief summary of the methods which were used during this research.

2.2.5.1 Cepstral mean subtraction

CMS was one of the earlier and most effective methods used for compensating cepstral features for linear channel induced effects in speech. In this method feature vectors are computed from the N cepstral vectors $\overrightarrow{c}_{y;i}$, $i = 1, ..., N$ (i denotes for frame time index) from a speech utterance $y(n)$ by subtracting the cepstral mean, or average of all N cepstral vectors, from each of the original cepstral vectors $\overrightarrow{c}_{y;i}$:

$$\overrightarrow{c}_{cms;i} = \overrightarrow{c}_{y;i} - \overrightarrow{c}_{y;avg} \quad , i = 1, ..., N \tag{2.3}$$

where $\overrightarrow{c}_{y;avg} = \frac{1}{N}\sum_{i=1}^{N} \overrightarrow{c}_{y;i}$. The principle behind this approach is based upon the behaviour of the cepstrum under convolutional distortions [73]. If we assume that most channel distortions are stationary (for example that caused by different microphones, telephone handsets and audio channels), or at least slowly time-varying, then the effect of the channel appears as convolutive noise in the time domain and hence becomes an additive constant in the log cepstral domain. Hence, subtracting the mean of each cepstral coefficient over the whole utterance removes the channel induced offset and any other stationary speech components.

2.2.5.2 Mean and variance normalization

Cepstral MVN [74] is the same as CMS, but in addition each feature coefficient is normalized by the estimated variance of that feature over the whole utterance. By subtracting the mean and dividing by the variance, the distribution of the features will have zero mean and unit variance. This is to remove different cepstral coefficient distributions due to variable channel distortions [1].

2.2.5.3 Feature warping

Feature warping is known as histogram equalization in the image processing literature, and in the speech processing literature as cumulative distribution mapping. The main aim of feature warping is to construct a more robust representation of the each cepstral feature distribution. Feature warping has been shown in [75] and [2] to significantly improve language and speaker recognition accuracy, respectively. This technique maps the distribution of feature vectors to a standardized distribution over a specified time interval. Using this method, speaker verification task demonstrated superior performance to those using other methods including CMS and MVN [76]. According to [75] feature warping is best applied to the MFCC. Although feature warping is applicable for any standard probability distribution, the best result is attained by using normal distribution as the target distribution [76].

Figure 2.2.4 shows the distribution of features before and after applying feature warping. In this example the target distribution which we want to warp the current distribution to, is a normal distribution. With the new warped value calculated for the cepstral feature in the centre of the window, a sliding rectangular window of size N is applied, the typical window size (N) is 3-seconds. The sliding window is advanced frame by frame and a new entry is calculated.

The warping is a non-linear transformation from the original feature, q, to a warped feature, m. This method is applied separately for each cepstral coefficient and it assumes that the features are independent. This approach uses cumulative distribution function (CDF) matching to make the features more robust to different channel and noise effects. CDF matching is performed over a sliding window. Only the central frame of the window is warped based on CDF matching. The features in a given window of the utterance are sorted in descending order. Suppose the central frame has a rank R (between 1 and N). Its corresponding CDF value is approximated as:

$$\Phi = \frac{N + \frac{1}{2} - R}{N} \tag{2.4}$$

The warped value m should satisfy:

$$\Phi = \int_{z=-\infty}^{m} h(z) d_z \tag{2.5}$$

where $h(z)$ is the probability density function of standard normal distribution. The value of m can be quickly found by lookup in a standard normal CDF table.

$$\int_{y=-\infty}^{q} \mathrm{f}(y)\, \mathrm{d}y = \int_{z=-\infty}^{m} \mathrm{h}(z)\, \mathrm{d}z \tag{2.6}$$

Equation 2.6 shows a direct mapping of a source cepstral feature q (with measured distribution $f(y)$) to the warped component m (with distribution of $h(z)$).

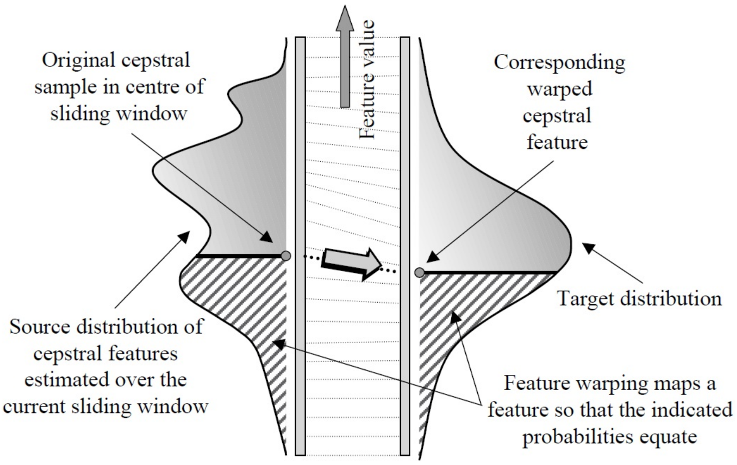

Original cepstral sample in centre of sliding window
Feature value
Corresponding warped cepstral feature
Target distribution
Source distribution of cepstral features estimated over the current sliding window
Feature warping maps a feature so that the indicated probabilities equate

Figure 2.2.4: Feature warping transformation (taken from [2]).

2.3 Modeling

2.3.1 Generative and Discriminative Classifiers

Generative classifiers acquire a model of the joint probability, $p(x, y)$, of the inputs x and the label y, and by using Bayes rules make their predictions to calculate $p(y|x)$ and then picking the most likely label y. But in discriminative classification the posterior $p(y|x)$ is directly modeled, or learned a direct map from inputs x to the class labels [77]. At this section two classifier types are compared.

In the area of automatic speaker characterisation, there are examples of employing generative, discriminative, and hybrid classifiers [78]. The classification stage is usually divided into two parts: *modelling* and *matching*. Based on the features extracted from a speaker's speech sample, a model is created of the speaker's voice, which is then entered into the identification system. Matching measures the similarity of the features extracted from an unknown speech sample with those of the speaker models [79]. The related speech data, known as training/enrolment data, are used to form a speaker-specific model. During the verification phase, the trained model is used to authenticate a sequence of feature vectors extracted from the utterances of unknown speakers.

Given a sequence of feature vectors from an unknown class (speaker, gender, age-group, or emotional state), the next task of the automatic speaker characterization is to classify that sequence as having come from one of the classes in the known population. Classification techniques can be divided into generative and discriminative (which are the two categories of the statistical approaches for constructing the models).

2.3.2 Statistical modelling (GMMs) and signal representation

Usually, in modelling-based approaches, the variable-duration speech signals are converted to the fixed-dimensional vectors, these vectors are then used by training/classification algorithms. To convert variable-duration speech signals into fixed-dimensional vectors, a probability density function (PDF) is fitted to acoustic features extracted from the speech signals such that the parameters of the fitted PDF characterise the speaker's identity, gender, age-group, and etc., then the fixed dimensional vector is formed by concatenating the mean vectors of the Gaussian Mixture Model (GMM). A GMM is used to model the complex distribution of acoustic features. Figure 2.3.1 shows the underlying idea of fitting a GMM to the acoustic features extracted from an utterance.

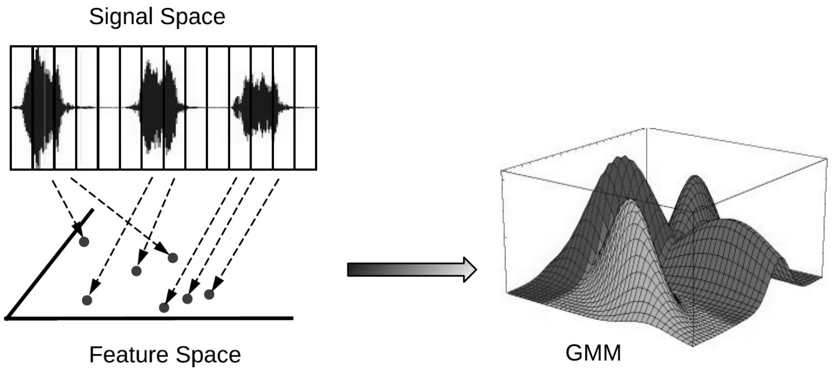


Figure 2.3.1: Graphical representation of feature extraction from speech signal and fitting a GMM to them (taken from [3]).

The amount of training data plays an important role in accurate estimation of the parameters of statistical models. For example, if the acoustic data include only the properties of a single utterance, fitting a GMM-based model to that utterance could not be performed accurately. This issue is more obvious in the case of using GMMs with a high number of mixture components. Therefore, methods used to adapt UBMs to characteristics of utterances typically draw upon the data in large training and testing databases.

2.3.3 Parameter estimation

The main proposed methods for parameter estimation are Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP).

2.3.3.1 Maximum likelihood estimation

In this method [80], given the class-dependent enrolment data, Gaussian means are estimated by maximising the likelihood of Equation 2.7.

$$\begin{gathered} p(x_t|\lambda) = \sum_{c=1}^{C} w_c p(x_t|\mu_c, \Sigma_c) \\ \lambda = \{w_c, \mu_c, \Sigma_c\}, c = 1, ..., C. \end{gathered} \tag{2.7}$$

where x_t is the acoustic vector at time t, w_c is the weight for the c_{th} mixture component, and $p(x_t|\mu_c, \Sigma_c)$ is a Gaussian probability density function with μ_c and Σ_c as its mean and covariance matrix, respectively. C is the total number of mixture components. This optimisation problem is challenging, as there is no information about the contribution of each sample to the mean of each Gaussian mixture component. Therefore, rather than directly maximising the log-likelihood of Equation 2.7, the auxiliary function of Equation 2.8, namely the complete data log-likelihood, is introduced and an iterative EM algorithm is applied [81]. During the E-steps of this technique, given the previous estimate of the model parameters , the auxiliary function in Equation 2.8 is formed by estimating the occupation counts $\gamma_{c,t}$ for each mixture component.

In the M-step, model parameters are updated by maximising the auxiliary function from the E-step. In [81], Bilmes showed that the maximisation of the auxiliary function over the model parameters increases the data likelihood of Equation 2.7. The new model is then considered as the initial model in the next iteration, and this iterative process is continued until convergence. The new model in each step is obtained by maximising the auxiliary function of Equation 2.8.

$$\Phi(\lambda, m_c) = \sum_{t=1}^{\mathcal{T}} \sum_{c=1}^{C} \gamma_{c,t} log[w_c P(x_t|m_c, \Sigma_c)], \tag{2.8}$$

where λ is the current model and m_c, Σ_c, and w_c are it's parameters. $\gamma_{c,t}$ is the occupation count for the c^{th} mixture component and the t^{th} segment. C and $\mathcal{T}$ are the total number of mixture components and time segments, respectively. Occupation counts are calculated as follows:

$$\gamma_{c,t} = \frac{w_c p(x_t|\mu_c, \Sigma_c)}{\sum_{c=1}^{C} w_c p(x_t|\mu_c, \Sigma_c)} \tag{2.9}$$

At the end, the means are calculated as follows:

$$m_c = \frac{\sum_{t=1}^{\mathcal{T}} x_t \gamma_{c,t}}{\sum_{t=1}^{\mathcal{T}} \gamma_{c,t}} \tag{2.10}$$

From Equation 2.10, a Gaussian mean will change with respect to its occupation count, which is related to the phonetic context that covered in the training utterance. Consequently, the MLE approach is not appropriate for modeling short utterances.

2.3.3.2 MAP adaptation

MAP is an approach to Gaussian mean adaptation [82]. This method involves a two-step estimation process similar to that of the MLE method. The first step of MAP and MLE are identical. In the second step of the MAP algorithm, the obtained sufficient statistics estimated in the first step are combined with the statistics of the prior mixture components parameters using the mixing coefficient η^{μ}, thus controlling the balance between the prior and the new information. In other words, this mixing coefficient controls the effect of new information on the mean parameter of the previously trained UBM.

A UBM could be defined by the following likelihood function:

$$\begin{aligned} p(x_t|\lambda) &= \sum_{c=1}^{C} w_c p(x_t|\mu_c, \Sigma_c) \\ \lambda &= \{w_c, \mu_c, \Sigma_c\}, c = 1, ..., C. \end{aligned} \tag{2.11}$$

where x_t is the acoustic vector at time t, w_c is the weight for the c_{th} mixture component, and $p(x_t|\mu_c, \Sigma_c)$ is a Gaussian probability density function with μ_c and Σ_c as its mean and covariance matrix, respectively. C is the total number of mixture components. The UBM parameters are estimated using a large amount of training data. The parameters of the adapted GMMs (w_c, μ_c, and Σ_c) are used for characterizing the utterances.

In MAP, the adapted means after the first iteration are estimated as follows:

$$m_c = \frac{\gamma_c m_c^* + \eta^{\mu} \mu_c}{\gamma_c + \eta^{\mu}} \tag{2.12}$$

where η^{μ} is the mixing coefficient and m_c^* is calculated as follow:

$$m_c^* = \frac{\sum_{t=1}^{T} x_t \gamma_{c,t}}{\sum_{t=1}^{T} \gamma_{c,t}} \tag{2.13}$$

$$\gamma_c = \sum_{t=1}^{T} \gamma_{c,t} \tag{2.14}$$

where $\gamma_{c,t}$ is the occupation count and could be found using Equation 2.9.

From Equation 2.12, it is obvious that mixture components c with high posterior probabilities mainly rely on the new adaptation data, but components with low posterior probabilities rely more on the information from the prior distribution.

2.4 Vector Space Representation

The high performance of Support Vector Machines (SVM)s in a range of technologies for speech detection, such as speaker detection, has been demonstrated beyond a doubt [83]. A nonlinear mapping between an input space and an SVM-expansion space (S-space), which is likely to be of high dimensionality, is carried out by SVMs. The kernel is the primary design element in an SVM. Internal products and distance metrics have a mutual effect on each other and therefore the identification of a suitable metric in the SVM feature space that can address the issue of classification is an important objective in SVM kernel design.

In recent times, the application of latent factor analysis to make up for speaker and channel variation has been an important research focus in GMM speaker recognition [84]. This approach involves the modelling of MAP-adapted means of a GMM with the use of latent factors for variability description. Moreover, the approach relies heavily on the use of a GMM super-vector made up of the stacked means related to the mixture elements. GMM channel compensation can be conducted by using this GMM supervector in conjunction with latent factor analysis.

In following subsections, these recent techniques will be discussed.

2.4.1 Support vector machines

SVMs are a group of related supervised learning techniques used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm makes a model that predicts the category into which a new example will fall. Naturally, an SVM model is a representation of the examples as points in space, mapped so that a clear gap, that is as wide as possible, divides the examples of the separate categories. Based on which side of the gap they fall, new examples are then mapped into that same space and predicted to belong to a category.

The hyperplane that has the largest distance to the nearest training data points of any class (the so-called functional margin) achieves a good separation, since in general the larger the margin, the lower the generalisation error of the classifier.

2.4.1.1 Theory of linearly separable binary classification

Suppose we have L training points, where each input X_i has D dimensions and is in one of two predefined classes $y_i = +1(or) - 1$, so our training data is of the form:

$$\{X_i, y_i\} \qquad where \qquad i = 1, ..., L \qquad and \qquad y_i \in \{-1, 1\}, \qquad X \in \Re^D$$

By assuming that the data are linearly separable, a line can be drawn on a graph of $X1 Vs. X2$, to separate the two classes when $D = 2$ and the hyperplane on graphs of $X_1, X_2, ..., X_D$ for $D > 2$.

In the general form, the hyperplane can be described by $w.X + b = 0$, where w is normal to the hyperplane and $\frac{b}{|w|}$ is the perpendicular distance from the hyperplane to the origin. Support vectors are the data points that are closest to the separating hyperplane. The aim of the SVM is to orientate this hyperplane in such a way as to be as far as possible from the closest support vectors of both classes.

As shown in the Figure 2.4.1, implementing an SVM boils down to selecting the variables w and b, so that our training data can be described by:

$$\begin{aligned} X_i.w + b &\geq +1 \qquad for \qquad y_i = +1 \\ X_i.w + b &\leq -1 \qquad for \qquad y_i = -1 \end{aligned} \tag{2.15}$$

And these two conditions can be combined together and be written as:

$$y_i(X_i.w + b) - 1 \geq 0 \qquad \forall_i \tag{2.16}$$

The support vectors are shown as circles in Figure 2.4.1. By considering these points (support vectors), the two planes on which these points lie can be described by:

$$\begin{aligned} X_i.w + b &= +1 \qquad for \qquad H_1 \\ X_i.w + b &= -1 \qquad for \qquad H_2 \end{aligned} \tag{2.17}$$

In order to orientate the hyperplane to be as far from the support vectors as possible, the margin needs to be maximized. Since the margin is proportional to $\frac{1}{||w||}$, maximizing the margin is same as minimizing $||w||$ such that $y_i(X_i.w + b) - 1 \geq 0 \qquad \forall_i$. For simplicity when we

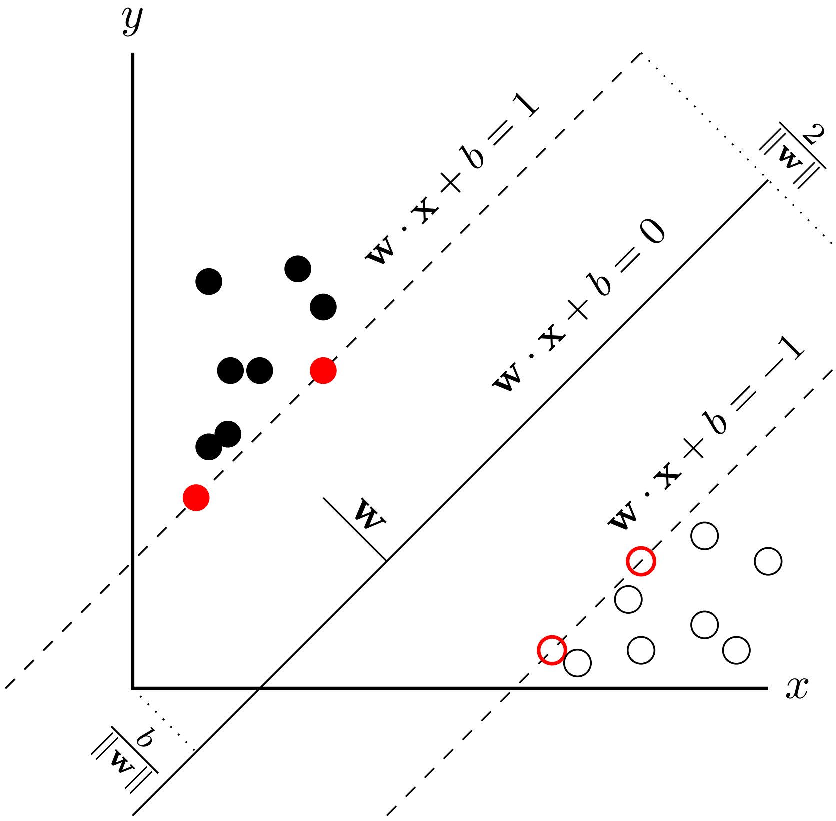

y
$\mathbf{w} \cdot \mathbf{x} + b = 1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = -1$
$\mathbf{w}$
x
$\frac{b}{\|\mathbf{w}\|}$

Figure 2.4.1: Hyperplane through two classes, which are assumed to be linearly separable.

apply quadratic programming (QP) optimization later we will try to minimize $\frac{1}{2}||w||^2$ which is equivalent to minimizing $||w||$.

Now, Lagrangian formulation of the problem needs to be presented. There are two reasons for doing this. The first is that the constraints (**??**) will be replaced by constraints on the Lagrange multipliers themselves, which will be much easier to handle. The second is that in this reformulation of the problem, the training data will only appear (in the actual training and test algorithms) in the form of dot products between vectors [85]. A Lagrange multiplier α is introduced, where $\alpha_i \geq 0$. This gives Lagrangian:

$$L_P \equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{L} \alpha_i y_i (X_i.w + b) + \sum_{i=1}^{L} \alpha_i \qquad (2.18)$$

This equation is generally distinguished as the prime form of the SVM.

Now we most minimize L_p with respect to w, b, and at the same time require that the derivatives of L_p with respect to all the a_i vanish, all subject to the constraints $a_i \geq 0$. Requiring that the gradient of L_p with respect to w and b vanish give the conditions:

$$w = \sum_{i=1}^{L} a_i y_i X_i \tag{2.19}$$

$$\sum_{i=1}^{L} a_i y_i = 0 \tag{2.20}$$

Equatin 2.19 points out that the vector w is a linear combination of the support vectors. Since these are equality constraints in the dual formulation, we can substitute them into Equation 2.18 to give

$$L_D = \sum_{i=1}^{L} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j X_i . X_j \tag{2.21}$$

As we see in the dual form, Lagrange multipliers and the new constraint for the optimum bias b should be kept. Only the dot product of each input vector X_i, which is another important notice in the dual form above (L_D), needs to be calculated. The only variable available in the dual form is the Lagrange multipliers a, so the aim now is to find these variables that maximise the dual form L_D, and keep the two constraints. Using the quadratic programming solver is the best way to solve this quadratic optimisation. w can be obtained by using the equation 2.19 once we get a, and then b by substituting w in: $y_s(X_s.w + b) = 1$, where x_s are the support vectors. All the parameters required to assign the maximal margin separating the hyperplane are w and b.

Generally, life is not simple, as described earlier. In particular, data are not linearly separable in speech applications, and so the linear SVM cannot be employed as described above. There are two methods to handle the non-linear separable data:

1. Introducing a slack factor that relaxes the constraints of the SVM slightly, by allowing for misclassified points, and then trading-off between the slack variable and the size of the margin.

2. Projecting the data into a high-dimensional space in which the data becomes linearly separable, and therefore a linear SVM can be applied.

The second method is more generally used in speech-related areas.

2.4.1.2 SVM-based speaker verification system

SVMs are supervised binary classifiers. When a kernel function is used, the optimal separator is given by sums of a kernel function $K(.,.)$,

$$f(x) = \sum_{i=1}^{L} a_i t_i K(x, x_i) + d \tag{2.22}$$

where the t_i are the ideal outputs, $\sum_{i=1}^{L} a_i t_i = 0$. The vectors x_i are support vectors that are obtained from the training set, L is the total number of training points, and d is the constant value. The ideal output is either 1 or -1. The vectors defining the hyperplanes can be chosen to be linear combinations with parameters a_i of images of feature vectors that occur in the data base. For classification, a class decision is based upon whether the value,$f(x)$, is above or below a threshold.

The kernel $K(.,.)$ is constrained to have certain properties (the Mercer condition [86]), so that $K(.,.)$ can be expressed as,

$$K(X, y) = b(X)^t b(y) \tag{2.23}$$

where $b(X)$ is a mapping from the input space (where X lives) to a possibly infinite-dimensional S space.

If we assume that the data set is separable, SVM chooses a hyperplane in the S-space with a maximum margin to separate the classes. The SVM training process models the boundary between classes. But in the real life they are not usually linearly separable, and even if they are, we might prefer a solution that better separates the bulk of the data while ignoring a few weird noise related behaviours. Please refer to Appendix B.1 for description of a distance measure algorithm and definition for a kernel function.

2.4.2 Factor analysis

Having data $x^{(i)} \in R^n$, which come form a mixture of several Gaussians, the EM algorithm can be applied to fit a mixture model.

In this setting, we usually imagine problems where we have sufficient data to be able to discern the multiple-Gaussian structure in the data. For example, this is true when the training size m (number of training-set vectors available) is significantly larger than the dimension n of the data. The dimensionality of training set vectors is equal to the multiplication of feature dimension with the total number of mixture components, so their dimension could be huge when a lot of components are used.

However, in a different scenario, where n is larger than m, data modelling is more challenging to achieve, regardless of whether single or multiple Gaussians are used. To be more exact, due to the fact that only a low-dimensional subspace of R^n is spanned by the m data points, modelling the data as Gaussian and applying standard maximum likelihood estimators to determine the mean and covariance would produce a singular covariance matrix. So if the the number of samples m is less than dimension of the data n, then the samples will be constrained in a proper subspace and the covariance matrix will be singular.

In the following sections, the factor analysis model is presented. Please refer to Section D.1 for description of several attributes of Gaussians employed later on, namely, the identification of marginal and conditional distributions of Gaussians.

2.4.2.1 The factor analysis model

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables. For example, it is possible that variations in four observed variables mainly reflect the variations in two unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus ”error” terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables. Factor analysis originated in psychometrics, and is used in behavioural sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data (this techniques is also applied for speaker recognition task [84]).

Factor analysis is related to principal component analysis (PCA), but the two are not identical. Latent variable models, including factor analysis, use regression modelling techniques to test hypotheses producing error terms, while PCA is a descriptive statistical technique. There has been significant controversy in the field over the equivalence or otherwise of the two techniques (please see Appendix F for more details).

For the factor analysis model, it is assumed that a joint distribution exists on (x, w), where

$w \in R^y$ represents a latent random variable:

$$\begin{aligned} w &\sim \mathcal{N}(0, I) \\ x|w &\sim \mathcal{N}(\mu + Tw, \Psi) \end{aligned} \tag{2.24}$$

here we have a super-vector space (S-space) with dimensions, R^s and a subspace (Y-space) of dimensions $y < s$, R^y, where the latent factor (w) lives. $T : R^y - > R^s$ is an $s \times y$ matrix.

The vector $\mu \in R^s$, the matrix $T \in R^{s \times y}$, and the diagonal matrix $\Psi \in R^{s \times s}$ are established as the model parameters. In general, s is given and y is chosen to be smaller than s. Figure 2.4.2, shows the graphical example for factor analysis approach, using $s = 2$ and $y = 1$. Figure 2.4.2 shows the typical samples of $w^{(i)}$ in a one-dimensional sub-space. Then these data-points are mapped to the two dimensional s space, by $\mu + Tw$. This model envisioning that X's inside each mono Gaussian circles are considered as original data points $x^{(i)}$. The working assumption is that the sampling of a y dimension mono Gaussian $w^{(i)}$ produces every data point $x^{(i)}$. The subsequent calculation $\mu + Tw^{(i)}$ enables the mapping of $x^{(i)}$ to a y-dimensional affine subspace. Afterwards, the addition of covariance Ψ noise to $\mu + Tw^{(i)}$ generates $x^{(i)}$. Along the same lines,

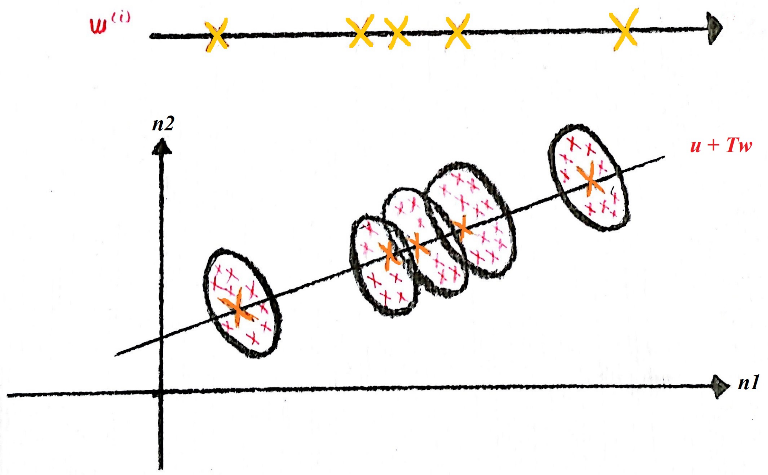


Figure 2.4.2: Graphical example of factor analysis based approach.

the factor analysis model can be defined based on

$$\begin{aligned} w &\sim \mathcal{N}(0, I) \\ \varepsilon &\sim \mathcal{N}(0, \Psi), \\ x &= \mu + Tw + \varepsilon \end{aligned} \tag{2.25}$$

where ε and w are independent [87].

To determine the precise distribution outlined by the model, it is important to note that the random variables w and x possess a joint Gaussian distribution

$$\begin{bmatrix} w \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{wx}, \Sigma). \tag{2.26}$$

Now μ_{wx} and Σ needs to be found.

As $w \sim \mathcal{N}(0, I)$, $E[w] = 0$. Also

$$\begin{aligned} E[x] &= E[\mu + Tw + \varepsilon] \\ &= \mu + TE[w] + E[\varepsilon] \\ &= \mu + T0 + 0 \\ &= \mu. \end{aligned} \tag{2.27}$$

Putting these together,

$$\mu_{wx} = \begin{bmatrix} 0 \\ \mu \end{bmatrix} \tag{2.28}$$

Next, to find Σ, we need to calculate $\Sigma_{ww} = E[(w - E[w])(w - E[w])^T]$ (the upper-left block of Σ), $\Sigma_{wx} = E[(w - E[w])(x - E[x])^T]$ (upper-right block), and $\Sigma_{xx} = E[(x - E[x])(x - E[x])^T]$ (lower-right block). Now, since $w \sim \mathcal{N}(0, I)$, we easily find that $\Sigma_{ww} = Cov(w) = I$. Also,

$$\begin{aligned} E[(w - E[w])(x - E[x])^T] &= E[w(\mu + Tw + \varepsilon - \mu)^T] \\ &= E[ww^T]T^T + E[w\varepsilon^T] \\ &= T^T. \end{aligned} \tag{2.29}$$

In the last step, we used the fact that $E[ww^T] = Cov(w)$ (since w has zero mean), and $E[w\varepsilon^T] = E[w]E[\varepsilon^T] = 0$ (since w and ε are independent, and hence the expectation of their product is the product of their expectations). Similarly, Σ_{xx} could be found as follows:

$$\begin{aligned} E[(x - E[x])(x - E[x])^T] &= E[(\mu + Tw + \varepsilon - \mu)(\mu + Tw + \varepsilon - \mu)^T] \\ &= E[Tww^T T^T + \varepsilon w^T T^T + Tw\varepsilon^T + \varepsilon\varepsilon^T] \\ &= TE[ww^T]T^T + E[\varepsilon\varepsilon^T] \\ &= TT^T + \Psi. \end{aligned} \tag{2.30}$$

By putting everything in its place we have:

$$\begin{bmatrix} w \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & T^T \\ T & TT^T + \Psi \end{bmatrix}\right). \tag{2.31}$$

Hence, we also see that the marginal distribution (please refer to Section D) of x is given by $x \sim \mathcal{N}(\mu, TT^T + \Psi)$. Thus, given a training set $x^{(i)}; i = 1, ..., m$, we can write down the log likelihood of the parameters [87]:

$$\mathcal{L}(\mu, T, \Psi) = log \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}|TT^T + \Psi|^{1/2}} exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T(TT^T + \Psi)^{-1}(x^{(i)} - \mu)\right) \tag{2.32}$$

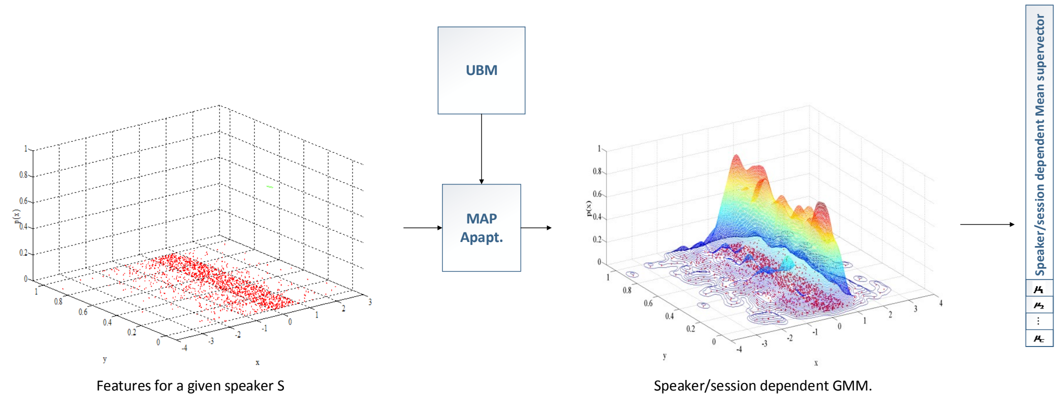
To perform maximum likelihood estimation, we would like to maximize this quantity with respect to the parameters. But maximizing this formula explicitly is hard, and we are aware of no algorithm that does so in closed-form. So, we will instead use to the EM algorithm. More details is included in Section 2.4.2.2.

2.4.2.2 I-vector

An initial attempt of applying factor analysis based approaches to speaker recognition problem made on 2005 by Kenny [84] (a brief description of this system is included in Appendix C).

More recently, the approach suggested by Dehak et al. [88] is based on defining only one space, as opposed to the multiple spaces defined in JFA approach. The new single space, which is called the total variability space, contains information about both speaker and channel variabilities. In this method mapping is performed using an $s \times y$ matrix, called T-matrix, from a low dimensional total variability space(Y-space) into the super-vector space (S-space): $T : R^y - > R^s$.

Given an utterance, the new speaker- and channel-dependent GMM super-vector is computed, as illustrated graphically in Figure 2.4.3. GMM mean super-vectors are obtained by MAP adaptation of UBM using a class dependent enrolment data and then the adapted means are concatenated to form a speaker/session dependent mean super-vector.



Features for a given speaker S

Speaker/session dependent GMM.

Figure 2.4.3: Graphical example of how to get the speaker/session dependent super-vector for a given speaker.

As it said earlier this new approach is based on defining only one space, as opposed to the multiple spaces defined in JFA approach. So in this method the obtained super-vector defined by rewriting Equation (C.1) as

$$x = \mu + Tw \tag{2.33}$$

where μ is the mean super-vector (concatenated means of UBM), T-matrix is the low-rank rectangular matrix, and w is the low-dimensional total variability vector, which is assumed to have a standard normal distribution $N(0, I)$. The training of the T-matrix is exactly the same as that for the eigenvoice V matrix in [89], except that during T-matrix training, all conversation sides of training speakers are treated as if they belong to different speakers.

In this framework, T and w are estimated using the EM algorithm. In the E-step, T is assumed to be known, and we update w. Similarly in the M-step, w is assumed to be known and we try to update T. The vector w is treated as a latent variable with the standard normal prior, and the i-vector is its MAP point estimate, which is obtained by maximisation of the following auxiliary function over w:

$$\Omega(\lambda, w) = \sum_{t=1}^{\mathcal{T}} \sum_{c=1}^{C} \gamma_{c,t} logw_c p(x_t | [\mu_c + T_c w], \Sigma_c) \mathcal{N}(w) \tag{2.34}$$

where $\mathcal{N}(w)$ is the distribution (normal) of w, and T_c is the row of the T matrix that corresponds to the mean of c^{th} Gaussian mixture.

In the E-step, the posterior distribution of w is Gaussian, with the following mean μ_w and covariance matrices σ_w [90]:

$$\sigma_w = [I + \sum_c \gamma_c T_c' \overline{\Sigma}_c^{-1} T_c]^{-1} \tag{2.35}$$

$$\mu_w = \sigma_w \sum_c [T' \overline{\Sigma}_c^{-1} \sum_t \gamma_{c,t} (x_t - m_c)], \tag{2.36}$$

in which I is the identity matrix of appropriate rank, and m_c and$\overline{\Sigma}_c$ are the adapted mean and covariance of the c^{th} Gaussian mixture. This set of parameters is updated during each EM iteration (for starting the algorithm, the UBM parameters are used).

In the M-step, the T matrix is estimated via maximisation of the following auxiliary function over T,

$$\widetilde{\Omega}(\lambda, T) = \sum_{s=1}^{S} \sum_{t=1}^{\mathcal{T}} \sum_{c=1}^{C} \gamma_{c,t,s} logw_{c,s} p(x_{t,s} | [\mu_s + T_c w_s], \Sigma_{c,s}). \tag{2.37}$$

Implementation steps towards obtaining i-vectors are described in Appendix E.

In Figure 2.4.4 a simplified block diagram of i-vector extraction and scoring is showed.

The front-end process is performed as described in Section 2.2.4 and feature warping applied on feature vectors using 3 seconds time window. This sequence of feature vectors is then represented by their distribution relative to a UBM, which is a GMM characterizing speaker-independent speech feature distributions. The parameters of this distribution are then transformed into an i-vector of R dimensions using a total variability matrix, T. As explained earlier T is a low dimensional subspace which contains factors of all variabilities (e.g. for speaker identification it contains factors of both speaker and channel variabilities).

After extracting the i-vectors, models are estimated using a discriminating projection. Ob-

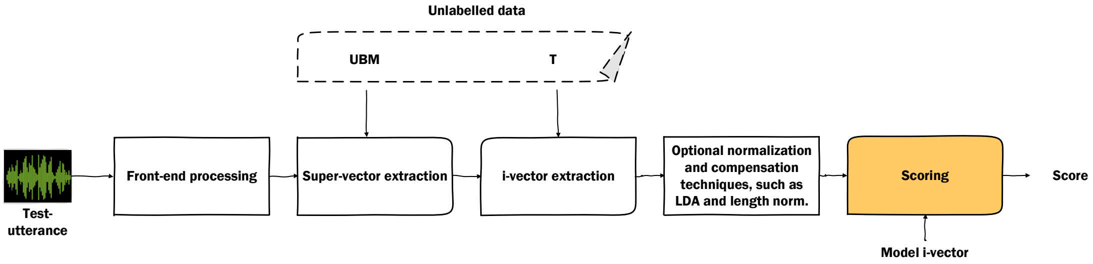


Figure 2.4.4: Simplified block diagram of i-vector extraction and scoring.

tained i-vectors represent all the variation in the speech in an utterance, but LDA is used to "pick out" which aspect of the speech we are attempting to discriminate e.g. speaker's identity, age, gender etc. Finally, a score between a model and test i-vector is computed. The simplest scoring function is the cosine distance between the i-vector representing the speaker's characteristic model (average of i-vectors from the training segments) and the i-vector representing the test segment. There are other methods for scoring which is explained in the following subsection.

2.4.2.3 Classification and scoring algorithms

Carrying out channel compensation in a low-dimensional total factor space, rather than in the GMM super-vector space, opens a new opportunity for assessing a variety of newly formulated channel compensation and scoring algorithms.

The choice of classifier greatly depends on the application. This selection could be influenced by the level of user cooperation, the expected channel, the amount of enrolment/test data, and the available computational/memory resources.

- **SVM classifier**

In 2009, a study by Dehak et al.[91] investigated the performance of an SVM as a classifier in the i-vector space (Y-space). In this study, i-vectors were used as a parameter input to SVM. The speaker factor coefficients correspond to speaker coordinates in the speaker space defined by the eigenvoice matrix. These vectors were tested with three classical kernels [91]: linear,

$$k(w_1, w_2) = \langle w_1, w_2 \rangle \tag{2.38}$$

Gaussian,

$$k(w_1, w_2) = exp(-\frac{1}{2\sigma^2}||w_1 - w_2||^2) \tag{2.39}$$

and cosine,

$$k(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{||w_1||\ ||w_2||} \tag{2.40}$$

kernels. Based on the Dehak et al. findings, the use of the cosine kernel function provides the best classification results [87], Equation 2.40. Where in this Equation, w_1 and w_2 are two extracted i-vectors from different utterances. Application of this cosine kernel function consists of normalising the linear kernel by the norm of both i-vectors. By removing the effect of magnitude, the focus is on the angle between the two i-vectors. 'It is believed that non-speaker information (such as session and channel) affects the i-vector magnitudes so removing magnitude greatly improves the robustness of the i-vector system' [87],

$$Cos(\theta) = \frac{\langle w_1, w_2 \rangle}{||w_1||\ ||w_2||} \tag{2.41}$$

where "θis the angle between w_1 and w_2.

- **Cosine Distance Scoring**

Another scoring algorithm which is proposed by Dehak et al. [87], was based on using the value of the cosine of the angle between the target speaker i-vector and the test i-vector as a decision score,

$$score(w_{target}, w_{test}) = \frac{\langle w_{target}, w_{test} \rangle}{||w_{target}||\ ||w_{test}||} \tag{2.42}$$

The value of this kernel is compared to the threshold for making a final decision.

The main advantages of this scoring over the SVM scoring are:

1. There is no need for enrolling the target speakers, as all of the speaker i-vectors are considered as a target speaker i-vector and at the end the one with the highest similarity will be picked as a true speaker.
2. Accepting the output of the cosine kernel as a decision score makes the process faster and less complex, compared to other scoring methods.

- **Scoring Method from Language Recognition Field**

All previously described methods were proved to be effective for speaker-recognition tasks. In 2012, Singer et al. [92] proposed a simple and effective scoring algorithm for language-identification tasks.

In this approach, linear discriminant analysis (LDA) [93] is used to find a new basis for the total variability space such that for any d, the subspace (D-space) spanned by the first d LDA basis vectors maximises the between-class variability while minimising the within-class variability. LDA is applied to the i-vectors for all training data from all classes and defines a projection matrix A^T of size $d \times y$ from the total variability space onto the D-space spanned by the first d LDA basis vectors. d is usually set to $Q - 1$, where Q is the number of classes.

The scoring is based on the dot product of the class model mean i-vector (m_l) and test i-vector, after LDA and unit normalisation $(\tilde{w}_{test})$.

$$score_l = \tilde{w}_{test}^T.m_l \tag{2.43}$$

This scoring is very similar to cosine scoring, but in this method, class mean, m_l, is estimated using unit-normalised i-vectors.

$$m_l = \frac{\sum_{j=1}^{N_l} \tilde{w}_j}{|| \sum_{j=1}^{N_l} \tilde{w}_j||} \tag{2.44}$$

where N_l is the number of utterances for each class l and $\tilde{w}$ are the unit-normalised LDA i-vectors and defined as,

$$\tilde{w} = \frac{A^T w}{||A^T w||} \tag{2.45}$$

The overall block diagram of a i-vector based identification system, using this scoring method, is illustrated in Figure 2.4.5. Blue and green parts of the diagram corresponds to the training and testing phase of the process, respectively. X_n represents feature sample from utterances and Y_n is a class label for utterances.

The use of i-vectors for automatic characterization systems has several distinct advantages over the GMM-SVM approach, description of this system is included in Section 2.4.1.2. For example, although off-line computation, for example training the T-matrix, of i-vector approach is complex and time consuming but the relatively low dimensionality of extracted i-vectors significantly reduces the on-line computational costs, for example channel compensation and scoring techniques, compared to a GMM super-vector system. Thus, the method lends itself to

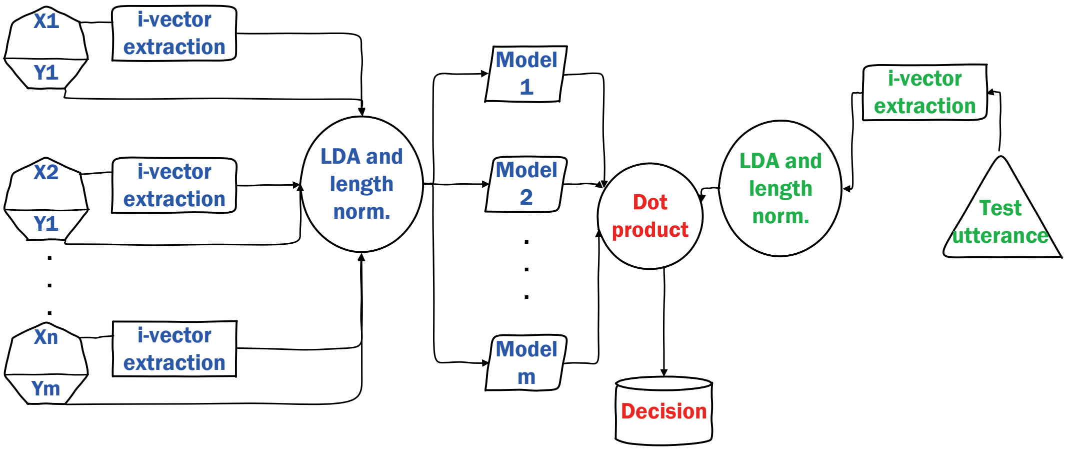


Figure 2.4.5: The block diagram of the automatic recognition system (which is designed for Age-ID) based on the i-vector approach, depicting both training (blue parts) and testing (green parts) phase. X_n and Y_n represent samples and class labels, respectively.

real-time implementation, which is important for applications.

The i-vectors are the low-dimensional total variability vectors, which are assumed to have normal distributions, and they are assumed to robustly represent an audio recording. I-vectors can be used as a new set of low-dimensional features for different classification purposes.

2.4.3 Inter Session Variability (ISV) modeling

One of the main factors that affect the performance of speaker characterization systems is variability, which is caused by changes in channel, speakers, and noise. There are many techniques in speech technology that are used to compensate for variability between channel conditions, such as feature normalization techniques and JFA, described in section 2.2.5 and 2.4.2, respectively. Another approach is inter-session variability (ISV) modelling, which is proposed as a session variability modelling approach that has been applied successfully to automatic speaker-characterization tasks [88, 94].

Session variability modelling aims to estimate and exclude the effects of within-class variation, in order to create more reliable class dependent models. The basic idea behind this technique is that the distortions due to ISV in the high-dimensional super-vector space can be summarized by a small number of parameters in a lower-dimensional subspace, which are called the channel factors [95]. The usage of this technique is examined on both the feature domain and the model domain.

One example of a modelling approach is in the Gaussian mixture model-universal background model (GMM-UBM) technique, where compensation is done by shifting the means of the world model and all of the class-dependent GMMs towards the ISV direction estimated from the test utterance, as shown in Equation 2.46.

$$\tilde{M} = \mu + Ux_c, \tag{2.46}$$

$\tilde{M}$ and μ are the compensated and original means of the UBM and all class-dependent GMMs in the super-vector domain. x_c is a D-dimensional vector representing channel factors for the test utterance. U (eigen-channel subspace matrix [88]) is a low-rank projection matrix for projecting the channel factors from low-dimensional ISV space to the high-dimensional super-vector space (S-space). The estimation of the eigen-channel subspace matrix (U-matrix) and the channel factors (x_c) is the same as the T-matrix and i-vector estimation, which has been explained in a Section 2.4.2.

In the i-vector approach we have a total variability space (Y-space), which summarises all variabilities. Then a classifier will focus on the relevant variability. But in ISV compensation we explicitly set up a space to describe the ”nuisance” variabilities.

2.5 Assessment of Detection Task Performance

Test trials for the automatic speaker characterization task can be categorized as either target trials or impostor trials. Each trial requires two outputs from the system under test. These are an actual decision, which declares whether or not the test segment contains the specified speaker's characteristic, and a likelihood score, which represents the system's degree of confidence in its actual decision. This can result in two types of actual decision errors, missed detections and false alarms. The miss rate ($P_{Miss|Target}$) is the percentage of target trials decided incorrectly. The false alarm rate ($P_{FA|Impostor}$) is the percentage of impostor trials decided incorrectly.

Test trials for the automatic speaker characteristics identification task can be categorized as either target trials or impostor trials. Each trial requires one output from the system under test, which is the actual decision. Speaker's characteristic model which obtains highest likelihood will be the out put of the identification system.

2.5.1 Evaluation measures for verification task

This speaker verification task NIST proposed the set of tools for assessing the performances of automatic recognition systems. Equal error rate (EER) is one of the intuitive measure and it is the miss (and false alarm) rate at the operating point where the two error rates are equal [96].

In addition to the single number measure of EER, more information can be shown in a graph plotting all the operating points of a system. An individual operating point corresponds to a likelihood threshold for separating actual decisions of true or false. By sweeping over all possible threshold values all possible system operating points are generated. This graph named Detection Error Trade-off (DET) curve by Martin et al. [97] as a assessment tool for detection task performance and it is became part of the 1996 NIST evaluation [98] for the representation of detection task performance.

For all verification experiments presented in this research, the obtained performances are represented by EER or/and DET curve.

2.6 Score Normalization Techniques

An important issue in the statistical approaches to speaker verification is score normalisation, which scales the log-likelihood score's distribution. Scaling the score distributions of different speakers is used to find a global speaker-independent threshold for the decision-making process. There are several commonly used score normalisation techniques, such as zero normalisation (Z-norm) and test normalisation (T-norm), which have been successfully applied to speaker-verification tasks [99–101].

The score normalisation techniques are mainly applicable during verification processes in which setting a threshold is strongly dependent on the distribution of impostor and true class scores. But in identification tasks, there is no need for normalising scores, as during identification the biggest score is chosen as a true class and the decision is not made based on the difference between true and impostor class scores.

2.6.1 Z-norm

In the Z-norm technique, the mean and standard deviation of the impostor scores are estimated off-line, and then the estimated mean is subtracted from each score, which is then divided by the estimated standard deviation, which allow us to use a global class-independent decision threshold:

$$\tilde{S}_s = \frac{S_s - \mu_i}{\sigma_i} \tag{2.47}$$

where, μ_i and σ_i are the estimated impostor parameters (mean and variance, respectively) for class model s. S_s is the log-likelihood score, and $\tilde{S}_s$ is the Z-normalized score.

2.6.2 T-norm

The T-norm also depends on mean and variance estimation for distribution scaling. During testing, a set of impostor models is used to estimate the impostor log-likelihood scores for a test utterance, and the mean and variance parameters are estimated from these scores.

Same as Z-norm, Equation 2.47 is also used for implementation of T-norm, but they are different in many ways. The main differences between the T-norm and Z-norm techniques are:

- The Z-norm technique tries to compensate for inter-speaker score variation, as against the T-norm technique, which compensates for inter-session score variations.
- The computed statistics are computed on-line in the T-norm technique, but these statistics are computed off-line in the Z-norm technique.

T-norm attempts to reduce the overlap between imposter and true score distributions of each class. Figure 2.6.1 shows the effect of T-norm score normalisation on distributions of impostor and true class scores (the effect on the mean only and not the variance). The dashed black and red arrows show the direction of the shift from the mean of the true (μ_T) and imposter (μ_I) score distributions before the application of T-norm, to the mean of the true ($\mu_{T'}$) and imposter ($\mu_{I'}$) score distributions after the application of T-norm. As Figure 2.6.1 shows, after applying

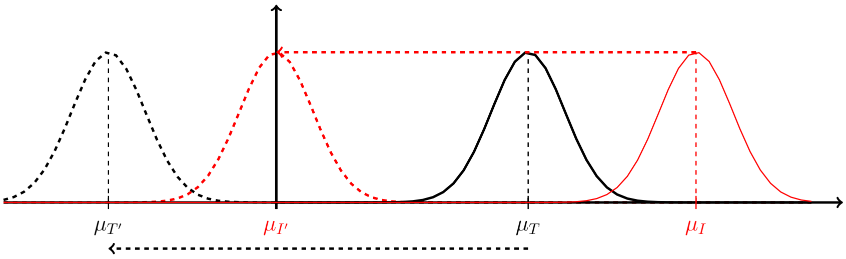


Figure 2.6.1: Effect of T-Norm on the distributions of true and impostor scores.

the t-norm on the scores, the means of the imposter class scores distribution will shift to the origin (zero) and the mean of the the true class scores distribution will move away from that.

2.6.3 Max-norm

Another method, which differs from the previously explained methods, is 'max-log-likelihood' score normalization. This method has been applied successfully to language ID systems [102]. In this normalization, the log-likelihood score of each target class is normalized with the score of the most competitive class model.

$$\tilde{S}_s = S_s - max_{i \neq s} S_i \tag{2.48}$$

In this method the mean of the true class score distribution will nearly move to zero and the imposter scores will move away from the origin, depend on their value (the smaller the scores, will results in a bigger shifts). This method, similar to most of the score normalization techniques, tries to separate out the true scores from the imposter scores.

2.7 The Performance of the Speaker Characterization Systems

This section outlines some of the applications, challenges and successful approaches for automatic speaker characterisation.

2.7.1 Related Works

During the last decades different approaches have been examined for automatic speaker characterization. First attempts to tackle this problem date back to the 1970s [103, 104]. Automatic identification of speaker characteristics approaches can be divided into phonotactic and acoustic approaches [105]. A phone recognizer followed by language models (PRLM) and parallel PRLM (PPRLM) techniques which have been developed initially for language recognition task, are successful phonotactic methods focusing on phone sequences as important information of different speaker characteristics such as language, accent/dialect, belonging to a particular social/regional group and even to an age category [106]. Phonotactic features and acoustic (spectral and/or prosodic) features provide complementary cues. State-of-the-art methods usually apply a combination of both through a fusion of their output scores [105].

Acoustic approaches are the main focus of this thesis. This approach does not need any specialized language knowledge [105]. Acoustic based approaches can be applied to identify paralinguistic speaker characteristics. They have been widely used in different speaker characterization problems [105, 107–111]. In [112–114], different types of acoustic features have been used with support vector machines (SVM) for speaker age-group identification. In [115], Gaussian mixture model (GMM) mean super-vectors and SVM were applied. In the field of speaker recognition, recent advances using i-vectors have increased the recognition accuracy considerably [87]. The same idea was also effectively applied to spoken language recognition [116]. Annual paralinguistic challenges held at INTERSPEECH provide a forum for state-of-the-art methods in speaker characterization such as emotional state and age recognition [111, 117]. In these challenges (year 2010 and specifically for age/gender and emotional state recognition), GMM mean super-vectors [118], GMM weight supervectors [119], Maximum-Mutual-Information (MMI) training [111], Joint Factor Analysis (JFA) [111] and fuzzy SVM modelling [120] have been suggested to enhance acoustic modelling quality.

In the automatic recognition of speaker characteristics, there usually exists a training data set, $D^{tr} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_n is the n^{th} utterance of the D^{tr}, and y_n is the class label of the utterance. This label corresponds to the speaker's characteristics. The goal is to find a function f, such that for an unobserved test utterance x^{tst}, output label $y^{tst} = f(x^{tst})$ is match the correct label. This problem could be approached using stochastic models or template-based models, both of which are classic approaches to automatic speaker characterization. The main focus of this research is on stochastic modelling approaches.

In stochastic models, speaker's characteristic features are modelled as a probabilistic source with an unknown but fixed probability density function. Over the training data, the parameters of the probability density function are estimated. The probability of the test utterance given speaker models are used for pattern matching. For text-independent and text-dependent speaker characterization, the most popular stochastic models are the Gaussian Mixture Model (GMM) [121, 122] and the hidden Markov model (HMM) [123, 124], respectively. In modelling approaches like GMM, the model parameters are obtained as the results of the estimation. Models like artificial neural networks (ANNs) [125] and SVM [126] model the boundary between speakers.

GMM is made up of multivariate Gaussian components [127]. A speaker's voice is characterized by a GMM super-vector of GMM parameters, such as the mean vectors, covariance matrices, and mixture weights. Using an expectation-maximization (EM) algorithm, the parameters of the model are typically estimated by maximum likelihood estimation [121, 122]. Because of its superior performance, probabilistic framework, and training methods scalable to large data sets, GMM is typically used in state-of-the-art speaker recognition systems [126].

Due to lack of class dependent training data for modeling the classes, Reynolds [100] introduced the GMM-UBM approach. In this approach, the UBM is trained from speech data collected from a large number of speakers, and acts as a speaker-independent model. Speaker models are obtained from the adaptation of a UBM through the maximum a posteriori (MAP) criterion [100]. The UBM is usually trained by means of an EM algorithm from a background data set that contains data on a wide range of speakers, languages, communication channels, and recording devices. Because of its reliable performance, the GMM-UBM has become a standard technique for identification of speaker characteristics.

The GMM-based approach also has disadvantages, one of which is that it models the features as a bag of frames, thus it ignores information about the sequence of for example, phones. Other modelling techniques such as HMMs have been explored by researchers to model sequential information of speech signals [128].

Later, the versatile classifiers known as SVMs have gained considerable reputation in the field of speech detection technologies [83]. An SVM is a discriminative classifier that models the boundary between a true and impostor classes. This technique finds a line or a hyperplane separating the two classes in the predefined kernel space.

Another approach has been to model blocks of features. It uses a mixed approach that combines the robustness of the statistical modelling provided by the GMM-UBM paradigm with the discriminating power of SVMs. A super-vector is extracted from the corresponding GMM (obtained from UBM by the MAP procedure), composed by concatenation of the mean coefficients of all the GMM components. The super-vectors are then used as inputs of the SVM classifier [129].

In recent years, joint factor analysis (JFA) [88] has emerged as a system that provided state-of-the-art performance for text-independent, speaker-recognition tasks in the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SREs) [90, 91]. It suggests a strong algorithm to specifically model inter-speaker variability and at the same time remove the channel or/and session variability. However, this method of modelling separate subspaces for capturing different speaker, channel, and session variabilities proved [130] to suffer from modelling of some speaker-specific information during channel factor training. Based on this fact, a study by Dehak et al. [131] proposed a new speaker-verification system based on factor analysis as a feature extractor. In this new approach, the concept of factor analysis is used to characterize a new low-dimensional space called the total variability space. In this space, a given speech utterance is represented by a low-dimensional vector named total-factor/i-vector. The main differences between the JFA and the i-vector approach are:

1. The new approach proposes defining only one space, compared to two spaces in JFA [88, 90].
2. Training the total variability matrix is the same as training the eigen voice [88] matrix, but all conversation sides of all training speakers are treated as belonging to different speakers.
3. The channel compensation in this new approach is carried out in the total variability space, compared to the high-dimensional GMM super-vector space in classical JFA [90].

2.7.2 Challenges in automatic speaker characterization

For identification of speaker characteristics, the recorded speech signal is the only available information, and there is no information about the articulatory system inputs, the physical states of the articulatory system, or the channel characteristics. Other technical restrictions that affect recognition performance are the amount and duration of speech recordings, the recording environment, the recording device, and the channel condition.

2.7.3 Automatic recognition of speaker's identity

While speech recognition aims to extract the underlying linguistic message in an utterance, speaker recognition aims to extract the identity of the person speaking the utterance. Since speech interaction with computers is becoming more pervasive, and its applications (such as telephone financial transactions and information retrieval from speech databases) are becoming more private and sensitive, there is a growth in the value of automatic recognition of a speaker based on vocal characteristics.

The successful usage of different types of classifiers and feature sets in the speaker-recognition task during past years is shown in Table 2.7.1.

Table 2.7.1: A Multidimensional Classification of Features in Speaker Recognition. EER stands for equal error rate, SRE for speaker recognition evaluation, LR for linear regression, GMM for Gaussian Mixture Modeling, and SVM for support vector machine.

Feature Type	Feature description	Model	Corpus	Performance	Ref.
Acoustic Feature (Base-Line)	Short-term cepstral-based	GMM	Switchboard-II	$EER = 3.3\%, 0.7\%$ for 1 & 8-conversation training, respectively	[100]
Prosodic features	Pitch and energy distributions	GMM	Switchboard-I	Using log-pitch+log energy+ their derivative; *EER* of 16.3%	[59]
	Pitch and energy track dynamics	Template	Switchboard-I	Using slop+duration achieved *EER* of 14.1%	[59]
	Prosodic statistics	LR	NIST's 2001	Using 19 statistics, duration & pitched related features; *EER* of 8.1%	[60]
Phone features	Phone *N*-grams	LR, SVM	Switchboard-I	Using "bag-of-*n*-grams" classifier achieved *EER* of 4.8%	[61]
	Phone binary trees	LR, binary tree	NIST SRE 2001	Using 3 token history (4-grams); *EER* of 3.3%	[62]
	Cross-stream phone modeling	LR, GMM	NIST SRE 2001	By fusing cross-stream & temporal system achieved *EER* of 3.6%	[63]
	Pronunciation modeling	LR	Switchboard-I NIST's 2001	By comparing word-level phone streams with open-loop streams; *EER* of 2.3%	[64]
Lexical features	Word *N*-grams	LR, SVM	NIST SRE 2004	Using n-gram idiolect system; *EER* of 9.0%	[65, 66]
Lexical-prosodic features	Duration-conditioned word *N*-grams	SVM	NIST's 2006 Switchboard-II	Using duration-conditioned lexical models; *EER* of 9.95%	[132]
Conversational features	Turn taking pattern & conversational style	GMM	NIST SRE 2001	Using conditional word usage results in; *EER* of 26%	[60]

2.7.4 Automatic identification of speaker's age and gender

In daily life, our voices are used to convey messages by words; however, our voices contain more than only the words we are saying. For example, due to its wide range of commercial/educational applications, such as interactive voice response systems, targeted advertising, and service customisation, speaker age estimation has been the subject of recent studies. However, automated speech-based age estimation is challenging for several reasons. First, there usually exists a difference between the perceived age of speakers and their actual age (or chronological age). Second, developing a robust age estimation method requires a database of speech from age-labelled speakers with a wide, yet balanced, range of ages. Third, speech contains significant intra-speaker variability that is not related to, or closely correlated with, age [133] including speaker weight, height, and emotional condition.

The Age-ID task attempts to predict the age of a speaker from a sample of his or her speech. This can be carried out in a classification scenario [134] using age groups or by using regression [135] (i.e., predicting the age in years). The use of GMM mean supervectors to model speech recordings prior to their implementation in support vector regression (SVR) is an efficient way to approximate age from speech [134]. A range of other problems of speech analysis, including speaker recognition, have been effectively addressed using similar SVM methods [87]. However, despite their efficiency, GMM mean super-vectors also present a major limitation in that due to their high dimensionality, they are not cost-efficient from a computational perspective, and moreover they make it difficult to develop a comprehensive model since data are restricted. To enhance the efficiency of GMM mean super-vectors in estimating age, PCA-based techniques have been employed to achieve dimension reduction [109, 136].

Perceived as a sub-element of speaker identification, gender identification may also be of significance when speaker identity is not of concern. For instance, numerous studies tend to address gender and age in association because perceptions of these traits have a direct effect on each other [115, 137]. By using a GMM mean super-vector and an SVM, Bocklet et al. created a number of seven age-gender groups to categorise speakers [115]. In addition, the recogniser used by these researchers encompassed MFFCs as features. Despite having some advantages, this technique requires the use of sizeable dimensions if there is a high number of Gaussians in GMM. In a more recent study [138], the use of an i-vector base system for age identification was studied. The advantage of this system compared to the system proposed by Bocklet, is that it does not require working with large dimensions, and it performs better than the old methods.

One of the main issues in identification of gender, specifically for teenage speakers, is the effect of rapid ageing due to their pubertal development. Rogol et al. [139] confirmed the presence of wide variation among individuals in the timing of the pubertal growth spurt and that a wide range of physiologic variations in normal growth was observed. The timing and tempo of puberty vary widely, even among healthy children, but there are several studies that estimate the distribution for age of pubertal growth for girls and boys, separately. For example, 11 years and 13 years were reported as average ages for the onset of puberty for girls and boys, respectively [139]. The same study claims that the growth during childhood is a relatively stable process [139].

Pubertal growth was defined as 'a dynamic period of development marked by rapid changes in body size, shape, and composition, all of which are sexually dimorphic' by [139]. Recently, the effects of puberty on the performance of gender identification were studied, and the authors confirmed that detecting the gender of a speaker during his/her pubertal growth is more difficult for both humans and machines [39].

The successful usage of different approaches in automatic speaker age and gender identification tasks during past years is shown in Table 2.7.2.

Table 2.7.2: Review of proposed automatic AGender recognition systems. In this table WSNMF stands for Weighted Supervised Non-Negative Matrix Factorization.

Task	Corpus	Methods	Number of classes	Performance	Ref.
AGender Rec.	German SpeechDat II Voice Class (eval.)	GMM-SVM	7 AGender classes	Precision=77% Recall=74%	[115]
AGender Rec.	German SpeechDat II Voice Class (eval.)	parallel phone recognizer	7 AGender classes	Precision=54% Recall=55%	[137]
		dynamic Bayesian networks	7 AGender classes	Precision=40% Recall=52%	
		linear prediction analysis	7 AGender classes	Precision=27% Recall=50%	
		GMM-UBM	7 AGender classes	Precision=42% Recall=46%	
Age Rec.	1:aGender 2:in-house(hebrew)	GMM-SVM (weight supervectors)	4 Age groups	1:Recall=59% 2:Recall=54%	[119]
AGender Rec.	aGender	GMM-UBM	7 AGender classes	ID rate 46%	[140]
		GMM-SVM	7 AGender classes	ID rate 43%	
		GMM-MLLR-SVM	7 AGender classes	ID rate 40%	
		Fused(acoustic and prosodic features)	7 AGender classes	ID rate 51%	
Gender Rec.	N-best evaluation	WSNMF	2 Gender classes	ID rate 96%	[133]

2.8 Summary

This chapter has presented a background review of the area of automatic speaker characterization form adult speech. Initially focused on the human production system, then some studies which attempted to correlate the perceptual cues that distinguish a person's voice to physical measurements of the speech waveform were discussed. A discussion of the features for automatic speaker characterization systems was presented. The attributes of ideal features were outlined along with the results of several feature selection studies indicating the use of spectral based features.

This was followed by an overview of the major modelling and classification techniques and their attributes used in speaker characterization systems. GMM-UBM, GMM-SVM, and two factor analysis based approaches (JFA and i-vector) are studied in this chapter. It is followed by a discussion on some of the successful classification algorithms, for i-vector approach, after which ISV compensation technique is explained.

Next three normalization techniques are studied and compared, z-norm, t-norm, and max-norm. At the end of this section some of the proposed methods for automatic speaker's identity, age, and gender identification tasks are presented.

"The more man realises his humanity, the lonelier he feels."

Ali Shariati

3

Speech Corpora

3.1 Introduction

Four different classification tasks are conducted in this thesis: speaker, accent, gender, and age recognition. The speech data used in training and evaluating the recognition systems are described in this chapter.

3.2 Kids Speech Corpora

3.2.1 OGI kids corpus

The OGI Kids' Speech corpus [141] is used to investigate performance of automatic speaker, gender and age-group identification systems. This corpus contains recordings of spontaneous and read speech, recorded at the Northwest Regional School District near Portland, Oregon. The corpus comprises of recordings of words and sentences from approximately 1100 children. A gender-balanced group of approximately 100 children per grade from kindergarten (5–6 year olds) through to grade 10 (15–16 year olds) participated in the collection. For each utterance, the text of the prompt was displayed on a screen, and a human recording of the prompt was played, in synchrony with facial animation using the animated 3D character 'Baldi'. The subject then repeated the prompt, which was recorded via a head-mounted microphone and digitized at 16 bits precision and 16 kHz sampling rate.

Table 3.2.1: Number of kids recorded for each grade.

Grade	Age	# of speakers	
	(years)	Male	Female
K	5-6	39	50
1	6-7	58	31
2	7-8	53	61
3	8-9	63	54
4	9-10	47	45
5	10-11	49	49
6	11-12	57	55
7	12-13	46	51
8	13-14	49	50
9	14-15	70	40
10	15-16	76	30

During this study, several experimental sets from the OGI data were used. Table 3.2.1 and Figure 3.2.1 show the number of children recorded per grade and the distribution of children's ages, respectively. In the first column of Table 3.2.1 the blue, red, and black grades correspond to the age group (AG), labelled as AG1, AG2, and AG3. Choosing a AG banding was a big challenge. It was a trade-off between amount of data available for train and test, and level of variability in each AG. We decided to give the first priority to the amount of available data for training and testing, as it will lead to more statistically reliable result. These age groups were used to investigate the problem of age and gender identification when using children's speech.

The divisions of these data by gender, age group, and speaker is illustrated in Figure 3.2.2.

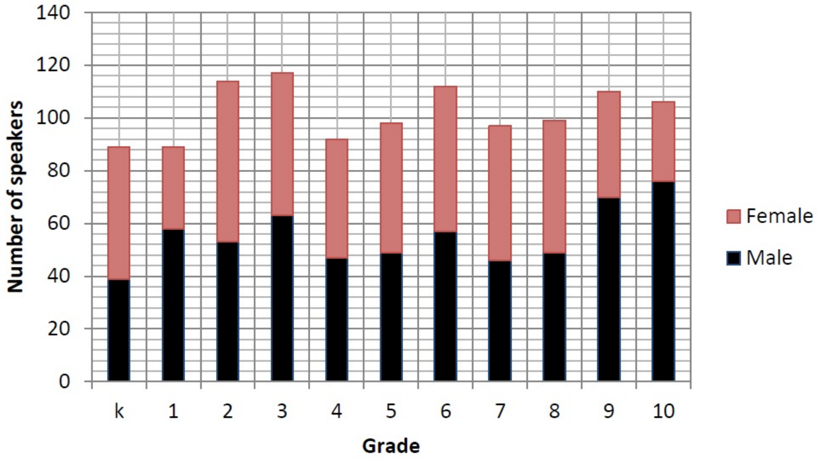


Figure 3.2.1: Distribution of ages of children recorded in the CSLU children's speech corpus.

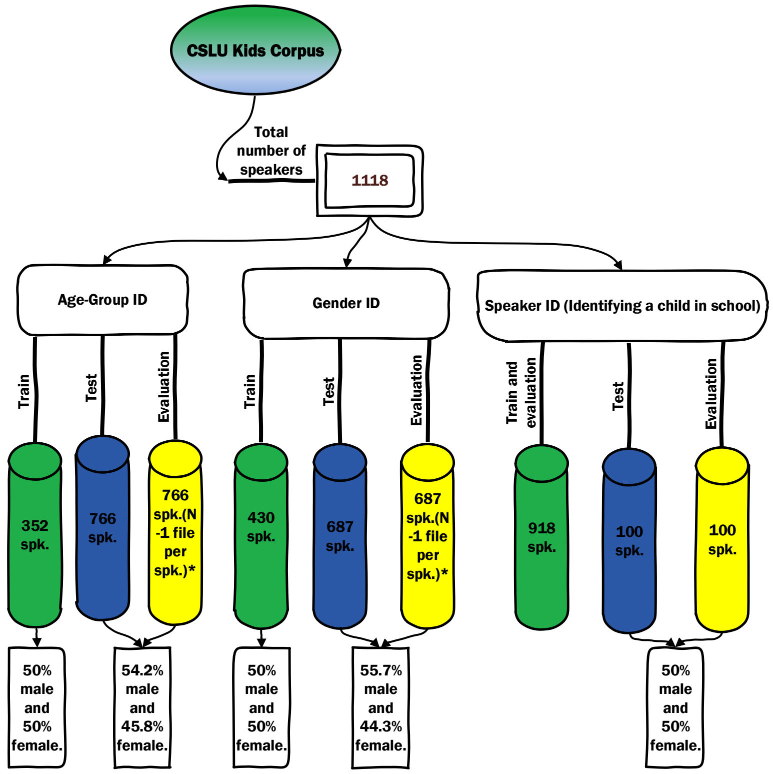


Figure 3.2.2: Partitioning data for age-group, gender, and speaker ID. *N is the total number of files per speaker, which varies from 30 to 69 for this corpus.

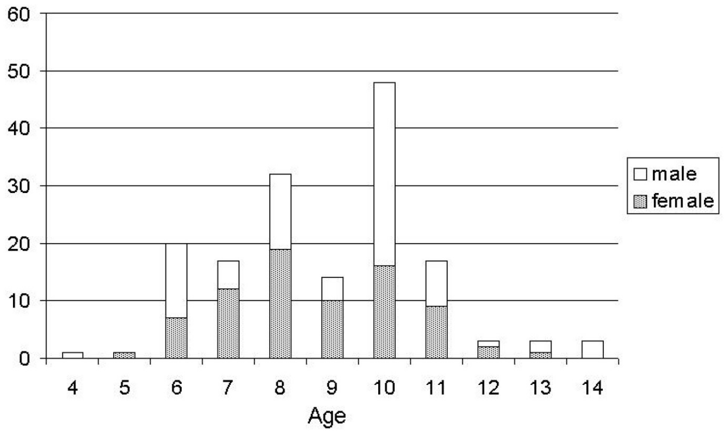


Figure 3.2.3: Distribution of ages of children recorded in the PF-STAR children's speech corpus [4].

3.2.2 PF-STAR corpus

The PF-STAR children's speech corpus [4] is used for the first evaluation of the utility of current speaker recognition techniques for children's speech. It comprises 14 hours of recordings from 158 British children (52% male), from Birmingham and Malvern, aged between 4–14 years, but with 92% of the children aged 6–11. The majority of the children (excluding some of the younger children) recorded 20 'SCRIBE' sentences, a list of 40 isolated words, a list of 10 'phonetically rich' sentences, 20 'generic phrases', an 'accent diagnostic' passage (the 'sailor passage') and a list of 20 digit triples. The recordings are divided into a training set (86 speakers, 703 recorded speech files, 7 hrs 29 mins 49 secs including non-speech), a evaluation set (12 speakers, 97 recorded speech files, 53 mins 58 secs including non-speech) and a test set (60 speakers randomly chosen, aged from 6-11 years old, 510 recorded speech files, 5 hrs 49 mins 47 secs including non-speech) [4] . The distribution of the children's ages is shown in Figure 3.2.3.

The speech was recorded at a 22.05kHz sample rate using close talking and desk microphones in a relatively quiet environment (typically a room or space off the school library), and recordings were made at three locations: a primary school in Malvern, Worcestershire (central England), a primary school in Birmingham, and an Industrial Assessment Centre (IAC) soundproofed booth in the department of Electronic, Electrical and Computer Engineering at the University of Birmingham. The texts were presented to the children on a laptop using in-house prompting and recording software.

3.3 Adult Speech Corpora

3.3.1 NIST 2003 data set

The NIST2003 corpus [142] is used to assess the performance of our baseline speaker recognition system (using telephony recorded speech signals) and compare our system with that proposed by Auckenthaler et al. [6, 101]. NIST2003 includes cellular data extracted from the Switchboard Cellular part 2. This corpus consists of 149 male and 207 female speakers with 2 minutes of training speech from a single cellular phone call. It contains just over 120 hours of English conversational telephone speech. A detailed description of the evaluation corpus is available on the NIST website [142] (In case of reading an electronic version of this thesis you may access this evaluation plan via: NIST-2003-SRE-Plan).

Details of the recorded data can be viewed in Table 3.3.1.

Table 3.3.1: details of recorded speech files, from NIST 2003 corpus.

Project	Sample		Applications	Data	Language
ID	Type	Rate		source	
NIST SRE 2003	8 bit u-law	8000	Speech and speaker recognition	Telephone conversation	English

3.3.1.1 Task Conditions

The NIST 2003 Speaker Recognition Evaluation (SRE) plan contained three different task conditions for the speaker detection task: one-speaker, detection-limited data; two-speaker, detection-limited data; and one-speaker; detection-extended data.

The focus of this research was on the first task condition: one-speaker, detection-limited data. This task condition is one of the most popular speaker detection tasks, and the conditions are:

- Only two minutes of training data from single conversation is available for a target speaker.
- Each test segment is a recording of the speech from the single speaker, and it should be only a minute from a single conversation.
- Non of the test materials should be used during the training process.

In the NIST plan, the verification task was defined as one against 10, indicating that at the scoring stage, each test utterance was scored against a true speaker model and the other 10 imposters. Finally, standard NIST software (DET-ware) was used to measure the verification performance [143].

3.3.2 TIMIT DATA SET

The TIMIT corpus [144] is used to assess the performance of our baseline speaker recognition system (using microphone recorded speech signals).

TIMIT contains a total of 6,300 sentences,10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. A speaker's dialect region is the geographical area of the U.S. where he or she lived during his or her childhood years.

Table 3.3.2: details of recorded speech files, from TIMIT corpus. PCM stands for Pulse Code Modulation.

Project	Sample		Applications	Data	Language
ID	Type	Rate		source	
TIMIT 1993	1-channel PCM	16000	Speech and speaker recognition	Microphone speech	English

The text material in the TIMIT prompts consists of 2 dialect 'shibboleth' sentences designed at SRI, 450 phonetically-compact sentences designed at Massachusetts Institute of Technology (MIT), and 1,890 phonetically-diverse sentences selected at TI. The dialect sentences (the SA sentences) were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences (the SX sentences), and each text was spoken by 7 different speakers. Details of the recorded data can be viewed in Table 3.3.2.

This database was used for the speaker-identification task. The divisions of the data are illustrated in Figure 3.3.1. TIMIT contains speech recordings from 192 female (30%) and 438 male (70%) speakers, from which 530 speakers were used for training of the background model and the remaining 100 (30 female and 70 male) speakers were used for tests. Unfortunately the database contains no information related to the age of speakers.

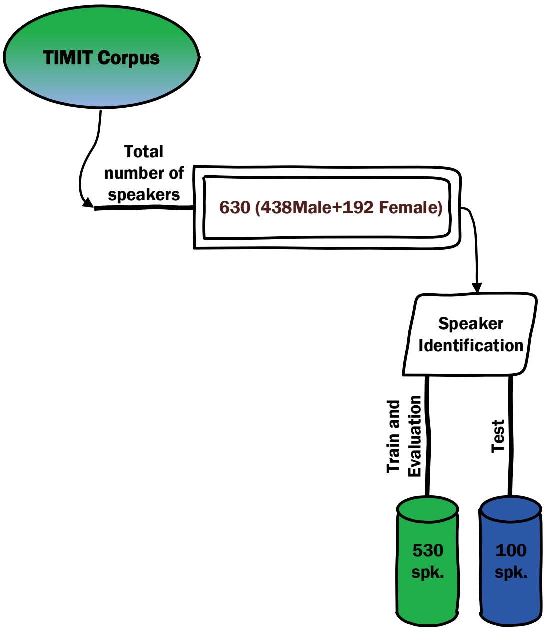


Figure 3.3.1: Partitioning data for speaker Identification.

3.3.3 The "Accents of the British Isles" (ABI) corpus

The ABI speech corpora were collected to support research into the implications of regional accents for speech and language technology. During this research ABI is used for investigation of the importance of different parts of the speech spectrum for speaker and accent identification.

The two ABI corpora comprise recordings of speech representing 26 regional accents of British English plus Standard Southern English (SSE). With the exception of SSE, all of the recordings were made on location in towns or cities that were judged to be representative of particular accents. The objective in each location was to record 20 subjects (10 men and 10 women) who were born in the location and had lived there for all of their lives. The SSE speakers were selected by a phonetician. Each subject recorded approximately 15 minutes of read speech. The prompt texts were chosen for their relevance to applications or their phonetic content. The microphones, recording and prompting software, and sample rate were the same as those employed in the compilation of the PF-STAR corpus. The recordings were made in relatively quiet rooms in libraries or community centres.

The ABI-1 [145]speech recordings represent 13 different regional accents of the British Isles, plus SSE, and they were used in all of the regional accent recognition experiments reported in Chapter 5. ABI-1 comprises recordings of 288 subjects: approximately/ideally twenty from each of 13 locations representing distinct accents of British English plus 20 subjects who were judged to speak SSE. ABI-1 consists of approximately 70 hours of recordings, with speakers' ages ranging from 16 to 79 years.

For the accent recognition experiments reported here, the head-mounted microphone recordings were bandpass-filtered (0.23-3.4 KHz) to simulate a telephone channel and down-sampled to 8.0KHz. Table 3.3.4 shows the details about the recorded speech files. The ABI-1 recordings are transcribed at the phrase level, but the transcriptions were not used in the present study. Table 3.3.3 shows the 14 regional accents form ABI-1 and their abbreviations that will be used throughout this thesis.

Accent	**Abbrev.**	**Accent**	**Abbrev.**
Birmingham	brm	Liverpool	lvp
Truro (Cornwall)	crn	Newcastle	ncl
Lowestoft (East Anglia)	ean	Denbigh (North Wales)	nwa
Hull (East Yorkshire)	eyk	Dublin (Republic of Ireland)	roi
Glasgow	gla	Elgin (Scottish Highlands)	shl
Inner London	ilo	Standard Southern English	sse
Burnley (Lancashire)	lan	Belfast (Ulster)	uls

Table 3.3.3: Accents of the ABI corpus and corresponding abbreviations

Table 3.3.4: details of recorded speech files, from ABI corpus.

Project	Sample		Applications	Data	Language
ID	Type	Rate		source	
ABI 2006	1-channel PCM	22050	Speech and speaker recognition	Microphone speech	English

ABI-2 was recorded using exactly the same methodology as ABI-1. It comprises approximately 70 hours of recordings of 286 speakers representing 13 regional accents of British English that were not covered in the original ABI-1 corpus. The material recorded is the same as in ABI-1, except that each subject recorded an additional set of 22 SCRIBE sentences. Aligned phrase-level transcriptions are not yet available for ABI-2, but were not required for the present study.

3.4 Summary

This chapter has introduced the speech corpora, for adult and children, that are used in the experiments described in this thesis.

"Raise your words, not voice. It is rain that grows flowers, not thunder."

Jalāl ad-Dīn Muhammad Rūmī

4

System Validation Experiments on Clean and Conversational Telephone Speech

4.1 Introduction

By using statistical modeling technique like GMM, classic speaker verification systems work to characterise the distribution of acoustic feature vectors that are observed while an individual talks. It is expected that this distribution contains information about the individual differences that are sufficient for identification of speakers. The distribution depends primarily on the physiology of the speaker's vocal tract, which governs the 'physics' of speech production, and the talker's 'idiolect', which is the particular sounds and sequences of sounds that the talker chooses to use.

The first intention of this chapter is to assess the performance of our baseline speaker recognition system and compare our system with that proposed by Auckenthaler et al. [6, 101]. The proposed system by Auckenthaler et.al uses a GMM-UBM speaker recognition system and they also investigate the usage of different score normalization techniques for speaker verification.

The obtained performances using different model sizes are presented in Figure 4.3.3. These performances will be then compared with our GMM-based baseline system for speaker verification task. The configuration of this base line system is included in Section 4.3.5.

In addition the performance of our i-vector based speaker identification system, which is obtained using microphone recorded speech signals, is presented in Section 4.4.

The aim of this chapter is not to compare the performance of different classification systems, but it is about parameter estimation and baseline validation of systems using both telephony and microphone recorded speech signals.

4.2 Data description

For testing the baseline systems, the TIMIT and NIST 2003 speech corpora were used. Full descriptions of these well-known corpora can be found in Section 3.3.2 and Section 3.3.1, respectively. During this research we needed to have two baseline systems, one for telephony recorded speech signals and one for microphone recorded speech signals (clean data).

The NIST 2003 corpus contains telephony recordings from adult speakers. Because published state of-the art results are available for this corpus, we initially used it to examine the performance of our classic GMM-UBM and GMM-SVM systems. In addition to the mentioned corpora, NIST 2002 was also used for training the UBM and calculation of score normalisation statistics. But none of the speakers from the NIST 2002 corpus appear in the test or evaluation sets of the experiments for this research.

The TIMIT corpus was used to evaluate the performance of the i-vector-based speaker recognition system. TIMIT is used as we wanted to use our i-vector based identification system for investigating different classification tasks using clean speech.

4.3 Speaker Verification System Based on GMM

Compared to other systems, the GMM speaker verification system has been proven to give good performance. Figure 4.3.1 shows the block diagram of the system design. All of our systems (GMM-UBM, GMM-SVM, and i-vector systems), from front-end to evaluation, are implemented using MATLAB.

As mentioned in Section 2.3.3, the EM algorithm [146] is one of the most popular algo-

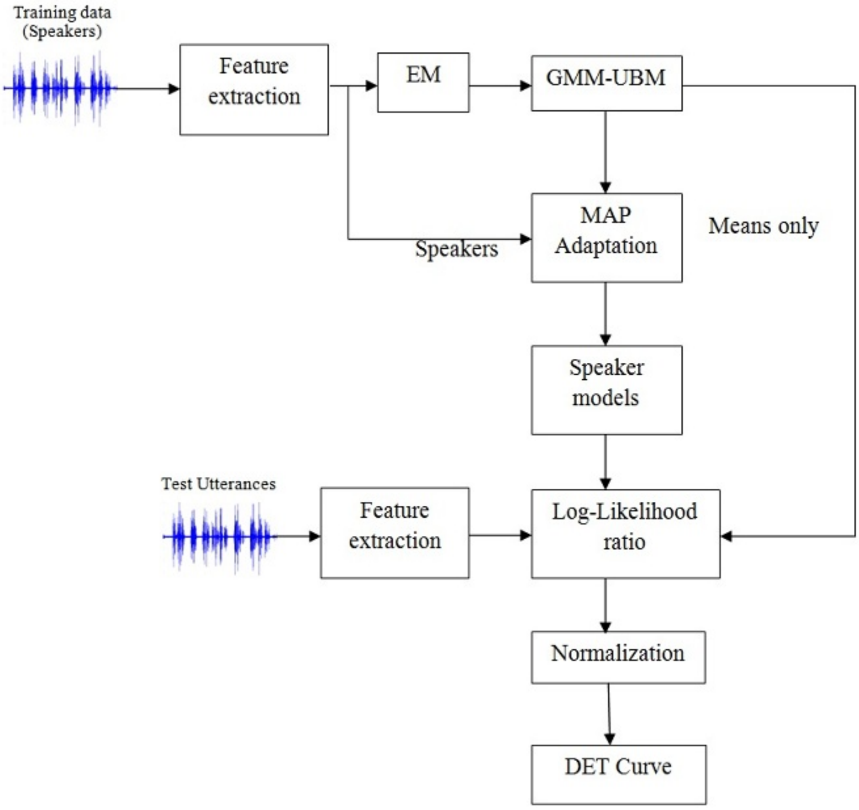


Figure 4.3.1: Block Diagram for the Speaker Verification System based on Gaussian Mixture Modeling

rithms for estimating the parameters of the Gaussian mixture PDFs. As described in Section 2.3.3, MAP adaptation was used to build the speaker dependent model, and 16 was used as relevance factor (mixing coefficient) in Equation 2.12 [100]. Only the means were adapted, as suggested in [100]. Adapting only means is computationally cheaper compared to mean plus weight and variance adaptation, but performances are at the same level and in some cases just slightly better when adapting all parameters instead of only means. But as the performance gain is small compared to the processing cost, we decided to update the means only and took the other parameters as specified in the UBM.

4.3.1 Feature extraction

The acoustic characteristics for speech recognition and for most of the automatic speaker characterization applications, can be captured by MFCC [147]. Please refer to Section 2.2.4 for more details. The first set of experiments are carried out with different numbers of coefficients and with different window size to determine the best configurations for feature extraction. The performance of the GMM based speaker recognition system using different configurations are presented in Table 4.3.1.

Table 4.3.1: Summary of the Basic Evaluation Results for Feature and Window Size Selection Experiments (for Speaker Verification using GMM-UBM system with 128 mixture components).

Experiment	Window		Number of coefficients		EER
ID	Size(ms)	Overlap(ms)	Static(cepstral)	Dynamic(delta)	(%)
1	32	16	19	19	**12**
2	32	16	12	12	12.5
3	20	10	19	19	12.3
4	20	10	12	12	12.8

The difference between MFCC order 19 and 12 seems not statistically significance, but we decided to use the MFCC order 19 for the rest of experiments.

4.3.2 Choosing dynamic features

As described in Section 2.2.3, for most speech classification tasks, the MFCCs are augmented with their delta (velocity) Δ, and double-delta (acceleration) ΔΔ, parameters [148], which provide local information about feature dynamics. For language ID, it was shown in [70] that improved performance can be achieved by incorporating broader temporal information using SDC coefficients. The description of this method is included in Section 2.2.4.2.

To choose between the proposed methods by [148] and [70], two sets of experiments were conducted, and the results are presented in Table 4.3.2.

Table 4.3.2: Comparison of dynamic feature sets for speaker ID using simple GMM-UBM system with 128 mixture components.

Experiment	Window		Number of coefficients		EER
ID	Size(ms)	Overlap(ms)	Static(cepstral)	Dynamic	(%)
1	32	16	19	19 (Delta)	**12**
2	32	16	19	49 (SDC)	25

Choosing 32ms window with 16ms overlap is based on the finding of the previous Section (please refer to Table 4.3.1). From Table 4.3.2, it is obvious that there is no improvement in the performance of the system that used SDC as a dynamic feature, compared to the delta coefficients.

4.3.3 Feature normalization techniques

As described in Section 2.2.5, many successful approaches have been proposed to normalize the features which were calculated from speech utterances. To compare FW and MVN (details included in Section 2.2.5.3 and 2.2.5.2, respectively), two experiments were conducted using the same configuration except for the feature normalization approaches. Table 4.3.3 shows the effect of using different feature normalization techniques on the performance of the speaker verification systems. Both systems are based on the GMM-UBM approach with a UBM of size 128 mixture components.

Table 4.3.3: Comparison of FW and MVN techniques, using GMM-UBM with 128 mixture components and 32ms windowing. Feature dimension is equal to 38*128=4864

Experiment	Number of coefficients		Normalization	EER
ID	Static(cepstral)	Dynamic(Delta)	technique	(%)
1	19	19	Feature Warping	**8.35**
2	19	19	MVN	10.5

The performance of the system when using the MVN technique was close to that of the system when using the FW technique, since both normalize the distribution of the features to have zero mean and unity variance. But the results confirmed that for the speaker verification task, FW works better than MVN, since FW forces the distribution of the features to be standard normal distribution with zero mean and unity variance. Mapping the raw features to a predetermined distribution, such as the standard normal distribution, appears to be a good way to make the features more robust to different channel and noise effects. MVN and FW are performed at utterance level. In order to make the features more robust to noise and channel effects, the FW approach was used as one of the enhancement techniques for the rest of experiments described in this chapter.

4.3.4 Effect of number of mixture components

In all previous experiments, we used the classic GMM-UBM system with 128 mixture components. This number was initially used to keep the computational cost low. However, we also wanted to examine the effect of the number of mixture components on the performance of speaker verification systems. To this end, a new set of experiments was designed using the best configurations as identified by previous experiments. The only parameter that was changed was the number of mixture components. Table 4.3.4 shows the results.

Table 4.3.4: Speaker verification performance in terms of EER, when using the full bandwidth and various numbers of mixture components. Feature dimension is the total number of mixture components * feature dimension(19+19).

Number of	Number of coefficients		Normalization	EER
mixture components	Static(cepstral)	Dynamic(Delta)	technique	(%)
1024	19	19	Feature Warping	6.15
2048	19	19	Feature Warping	**5.80**
4096	19	19	Feature Warping	5.90

The table shows that by increasing the number of mixture components, the performance of the system improved up to the saturation point. This saturation point depends on different factors, but it mainly depends on the total amount of training data available for the classification task. It can be seen that the best performance for adults is 5.80% EER with 2,048 mixture components. But the difference between the achieved performance by using 1,024 and 2,048 mixture components is negligibly small, and the use of 2,048 mixture components is expensive in terms of computational costs. The claim about the negligibly small improvement is based on the result of the McNemar's test [149] on the scores from system with 1024 (named classifier 0) and 2048 (classifier 1) mixture components GMM. The inputs to the McNemar's test are the counts of occasions when classifier 0 was correct and classifier 1 was incorrect ($= N_{01}$) and vice versa ($= N_{10}$). The result of this test for probability testing on the null hypothesis was $p = 0.4076$, which is not the evidence for a statistically significant performance improvement in the system which used bigger model size.

4.3.5 Base line system for telephony speech

The best performance was achieved by using the GMM-UBM speaker verification system with 2,048 mixture components, which has been used for the 2003 SRE conducted by NIST. For obtaining performance comparable to that of the state-of-the-art speaker recognition systems, our baseline system used the following settings:

- From the NIST-2003 speaker recognition evaluation plan, the one-speaker detection (limited data) task was followed (full description of task condition is available from Section 3.3.1.1).
- Speaker training data: One session was held for each speaker, and each session comprised about 2 minutes of speech from a single conversation.
- Test utterances: The speech duration was between 15 and 45 seconds.
- Front-end process: A filter-bank front-end with a 38-dimension feature vector, which comprises 19 MFCC and 19 delta coefficients, was used.
- Feature warping were applied as main feature normalisation method; it only applied on the static features.
- Speaker adaptation: Mean parameters only; variances and weights were identical to those used in the world model parameters.
- Score normalisation: Maximum normalisation was used; the full description of this technique is accessible from Section 2.6.3.
- For building the UBM the NIST-2002 database was used.

Figure 4.3.2 and Figure 4.3.3 show the performance of our baseline system and the performance of the system proposed by Auckenthaler et al., respectively.

As mentioned earlier, the main goal initially was to build a state-of-the art speaker verification system that would provide similar or even better performance than that provided by the proposed system by Auckenthaler et al. [6]. Their work is described fully in [6], and the DET curve obtained from their experiments is illustrated in Figure 4.3.3. The proposed system by Auckenthaler et al. used the following settings for the verification process:

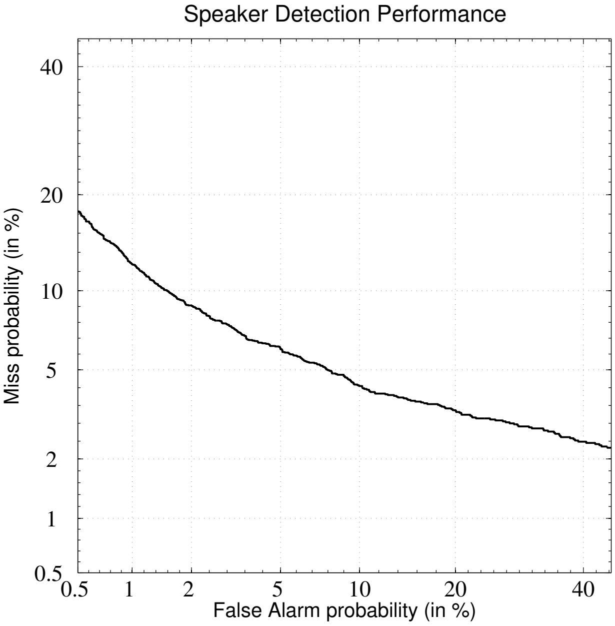


Figure 4.3.2: DET curves related to the performance of the baseline verification system.

- NIST 1998 evaluation database was used for experiments,
- Two sessions, each of a minute duration, were used for training speaker dependent models,
- Approximately ten seconds of speech was used for testing,
- Filter-bank front-end with 39 dimensional feature vector was used for converting speech into feature vectors,
- EM-algorithm was used for training of four hours of speech data for each gender, for creating two gender dependent world models.

These DET curves show that the best Auckenthaler et al.'s system had an EER of 7.00%, while our system had an EER of 5.80%. This difference could be due the technique which was used for feature normalization, in our system feature warping is used and in Auckenthaler et al.'s system MVN technique is used.

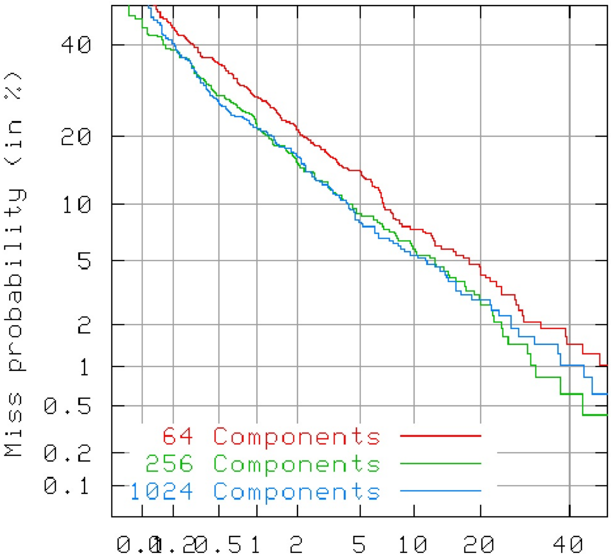


Figure 4.3.3: Performance of Auckenthaler et al.'s system (taken from [5, 6]).

4.4 Speaker Verification System Using i-vector

In addition to our baseline classic GMM-UBM system, we also designed a baseline system using the most recent configuration of state-of-the-art speaker recognition systems. The main idea of this recent method was proposed; namely i-vector, by Kenny et al. in [84, 89].

In this section, the results of evaluation experiments will be presented. The TIMIT speech corpus is used for all the experiments presented in this section. Full details of this corpus are given in Section 3.3.2. NIST database is not used for evaluation of our i-vector baseline system, as the i-vector frame-work is used only for microphone recorded data (clean speech), but NIST database contains recordings from telephone conversions.

For all the experiments in this section, a relatively small-scale task was designed using speech material from the TIMIT corpus, which contains recordings from 630 speakers. For background model training, we used all (there were 10 short sentences per speaker) sentences from each of 530 speakers (5,300 sentences in total).

100 speakers are used for testing and for speaker-specific model training, 9 out of 10 sentences per speaker were used, and the remaining 1 sentence was kept as a test. Verification trials were conducted with all possible model–test combinations, making a total of 10,000 trials (100 target vs. 9,900 impostor trials).

The speaker recognition system was implemented using MATLAB. The Microsoft Conversational Systems Research Centre, specifically Seyed Omid Sajadi, provided the MSR toolbox [150]. During this research, some functions from this toolbox were modified and used along with other self-written code.

The block diagram of our baseline i-vector system is depicted in Figure 4.4.1. The theory of all of the components of this diagram is presented in Section 2.4.2.

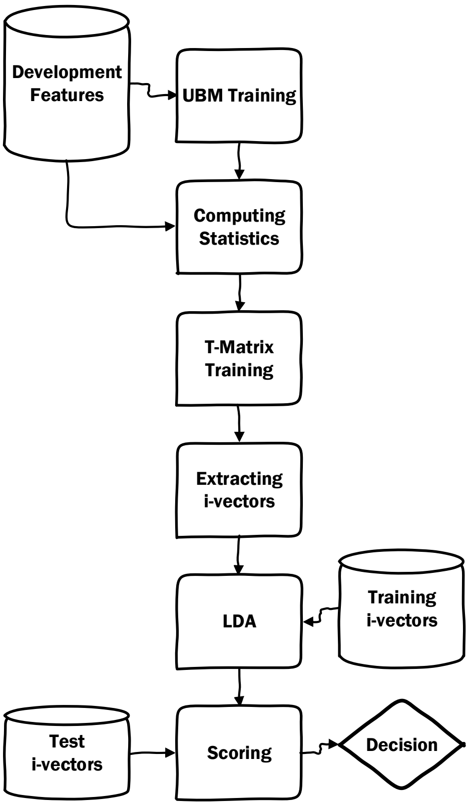


Figure 4.4.1: Block diagram of the i-vector system.

In the i-vector approach, i-vectors are the low-dimensional representations of an audio recording and they can be used for classification and estimation purposes.

The scoring approach proposed in [151] for language identification is used in this experiment (this approach is described in Section 2.4.2.3). In this approach the decision will be made based on the dot product of the unit-normalised LDA test i-vector with the class model mean, details included in Section 2.4.2.3 (Equation 2.43 is used for scoring).

4.4.1 Learning a total variability subspace from the observations

Prior to training the total variability subspace (T-matrix) we needed to have a trained UBM and calculated Baum-Welch statistic from development utterances, as explained in Section 2.4.2.

The EM algorithm was used for training the UBM with 256 mixture components. A 256-mixture-component UBM model was used based on the experimental results presented in Table 4.4.1. Based on the experiments on the evaluation set, the best results were obtained by using T-matrix with 400 dimensions, so this dimension is used for the rest of the experiments presented in this chapter. The results on the TIMIT data are an order of magnitude better than the NIST results, the reason is that TIMIT is the collection of clean speech recordings, as against with NIST which is the collection of telephony recorded speech.

Table 4.4.1: Speaker verification performance in terms of equal error rate (EER), when using the full bandwidth and various numbers of mixture components.

Number of mixture components	Number of coefficients		Scoring technique	EER (%)
	Static	Dynamic		
32	19	19	Simple doc product	3.20
64	19	19	LDA and dot product	2.00
128	19	19	LDA and dot product	0.95
256	**19**	**19**	**LDA and dot product**	**0.65**
512	19	19	LDA and dot product	0.74

Figure 4.4.2 and 4.4.3 show the DET curve and 3D-confusion matrix, respectively. Both these figures are related to the experiment which appears in boldface type in Table 4.4.1. Figure 4.4.2 is looks like steps as the test set is small. The usage of small test and training sets was due to the lack of computational resources. Figure 4.4.3 shows the effectiveness of i-vector method in separating out true speaker score from the imposter scores for each test utterance, as the diagonal (corresponds to true speaker scores) scores, in almost all cases, are much bigger than imposter scores.

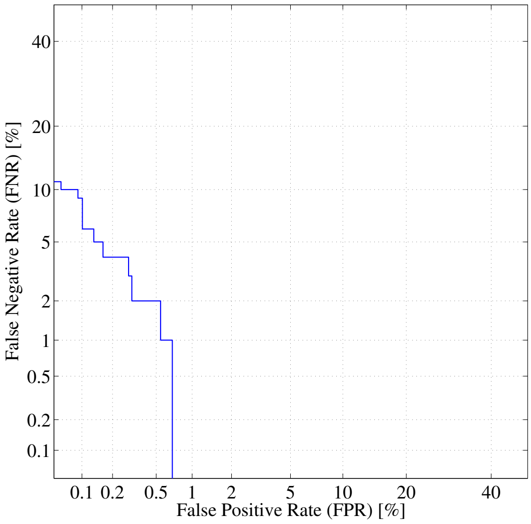


Figure 4.4.2: Det-curve for speaker identification using the i-vector system.

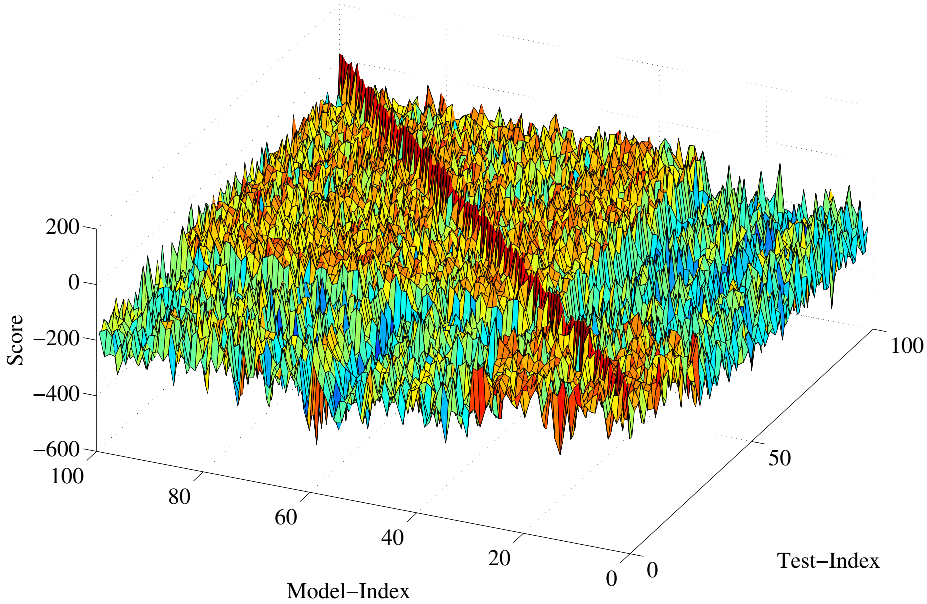


Figure 4.4.3: 3D-confusion matrix for the i-vector system.

4.5 Summary

The first step in this project was to develop a state-of-art speaker verification system. The NIST 2003 database was used along with techniques that were studied to reduce the environment mismatch between the training and testing conversations. As it has been shown in this chapter, based on the evaluative experiments, the baseline systems were designed, and the best obtained performance was compared with the performance of a state-of-the-art (at the time) speaker verification system. This comparison confirmed that the performance of our baseline system was at the same level and even better than the state-of-the-art speaker recognition systems (Auckenthaler et al.'s system). These experiments were first implemented using the classic GMM-UBM system, but based on the more recent proposed approach, the i-vector framework, a set of evaluative experiments was conducted and the effect of using different parameter sets on the overall accuracy of the system was also studied. The results show the performance of GMM-UBM and i-vector systems for speaker verification tasks (using different parameter sets), when using telephone conversational and microphone recorded speech, respectively.

This baseline system could be used for other types of automatic speaker characterisations, from both adult and child speech. The main area of interest for this research is the study of detection technologies from children's speech, as there are many differences between adult and children's speech signals but relatively little research has been done on children's speech compared to that from adult.

"Life is a balance of holding on and letting go."

Jalāl ad-Dīn Muhammad Balkhī

5

Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification

5.1 Introduction

We all know that an acoustic speech signal has information beyond its linguistic content. This paralinguistic information includes clues to speaker's accent and identity which automatic Accent IDentification (AID) and Speaker IDentification (SID) systems exploit them. The relationship between AID and SID is unequal, since accent information is related to SID but speaker information is a distraction in the context of AID.

Currently, for both AID and SID, the most commonly used parametrization is to represent a spoken utterance as a sequence of MFCC vectors derived from spectra, covering the entire frequency bandwidth. However, we know that different frequency regions contain different types of information. For instance, performed on the clean TIMIT corpus using mono Gaussian modelling, the SID study in [31], showed that the frequency regions below 600 Hz and above 3000 Hz provided better SID accuracy than the middle-frequency regions. However no similar study has been shown for AID. Using contemporary GMM-based systems, the contrasting importance of different frequency bands for AID and SID are investigated in this chapter.

5.2 Related works

In the most widely used approach to AID and SID, the distributions of feature vectors are characterized using a GMM [80, 152], as described in Section2.3. MAP adaptation of a UBM typically builds Individual accent or speaker GMMs. By using data from a variety of accents, speakers and background conditions speaker independent GMM is constructed. For various SID tasks [153], this approach has been very effective and its performance remains comparable to that obtained with more complex models. It has also been applied to AID, but with less achievement [105]. Using a discrimination-based approach, such as a SVM, applied to GMM super-vectors, which consist of the 'stacked' means of the mixture components of the accent or speaker GMMs [83, 105] is an alternative, please refer to Section 2.4.1.2 for more details. In [154], the GMM was used to calculate likelihood values and the SVM classifier was used to separate the likelihood values for a target speaker and impostor. The use of phone durations and average cepstra [155], phone and word-level HMMs [156–159], and stochastic trajectory models [160] are incorporated by other acoustic based approaches. The most successful systems use ISV Compensation, to remove irrelevant variability in speech classification tasks, this is a subspace projection technique which has been shown to improve the performance of speaker, language and accent identification and has become a standard component of these systems (Please refer to Section 2.4.3 for more details) [95, 105]. 'Phonotactic' approaches to AID exploit accent-dependent differences in the sequences in which speech sounds occur [161]. For AID, these approaches perform better than the GMM-based acoustic methods described above [105].

5.3 Data Description

In all experiments in this chapter the ABI-1 corpus of regionally accented adult's speech was used. This was collected to support research into the implications of regional accents for speech and language technology. The full description of this corpus is available in Section 3.3.3.

The speakers were divided into three subsets for both SID and AID; two with 93 and one with 94 speakers. In each subset gender and accent were distributed equally. A "jack-knife" training procedure was used in which two subsets were used for training and the remaining subset for testing. With different training and test sets, this procedure was repeated three times so that each ABI-1 speaker was used for testing, and no speaker appeared concurrently in the training and test sets. Using 993 segments of length 3, 10, and 30 seconds from all test recordings the SID systems were evaluated. The AID systems were evaluated using 1504 30-seconds segments from all test recordings. All the above mentioned segment lengths are after silence removal. For example for AID if speaker *s* is appeared in jack set 1 for test, then it will not appear for training the UBM and accent dependent models for experiment on jack 1, and it will appear for training purposes during experiments on jack 2 and jack 3. And for SID, if we have speaker *s* in jack set 1 for testing, then it will not be used for training the UBM, but for speaker *s* we have multiple files, one is used for test and the rest for estimating the parameters of speaker dependent model.

5.4 System Description

5.4.1 Signal analysis

For both SID and AID, feature extraction was implemented as follows. Using an energy-based SAD (application of pitch based SAD is also investigated for one experiment on AID), periods of silence were discarded. The speech was then segmented into 20-ms frames (10-ms overlap) and a Hamming window was applied. Obtained by applying the FFT, the short-time magnitude spectrum is passed to a bank of 31 Mel-spaced triangular band-pass filters, spanning the frequency region from 0 Hz to 11025 Hz. Table 5.4.1 shows the center frequency for each filter bandwidths, and Table 5.4.2 shows the frequency region which is covered by each sub-band.

At first the SID and AID experiments were performed using the full bandwidth (0–11.025 kHz) and telephone bandwidth (0.23–3.4 kHz) speech. By passing the recordings through a band pass filter, the latter was obtained. The calculation of MFCCs was based on all 31 filters and the first 23 filters for full and telephone bandwidth, respectively. In both cases, the first 19 MFCCs were used.

Using frequency band limited speech data comprising the outputs of groups of four adjacent filters (please refer to Section 1.4 for more details), separate SID and AID experiments were conducted to investigate the effect of different frequency regions. 28 overlapping sub-bands were considered, where the Nth sub-band comprises the outputs of filters N to $N + 3$ ($N =$

Table 5.4.1: The Center Frequencies for 31 Mel-spaced Band-Pass Filters

FILTER NUMBER	CENTER FREQUENCY (Hz)	FILTER NUMBER	CENTER FREQUENCY (Hz)
1	129	17	2239
2	258	18	2497
3	344	19	2799
4	473	20	3100
5	559	21	3445
6	645	22	3832
7	775	23	4263
8	861	24	4737
9	990	25	5254
10	1076	26	5857
11	1205	27	6503
12	1335	28	7235
13	1464	29	8053
14	1636	30	8957
15	1808	31	9948
16	2024		

Table 5.4.2: Map from sub-bands to the frequencies.

SUB-BAND NUMBER	FREQUENCY REGION (Hz)	SUB-BAND NUMBER	FREQUENCY REGION (Hz)
1	0-559	15	1636-2799
2	129-645	16	1808-3100
3	258-775	17	2024-3445
4	344-861	18	2239-3832
5	473-990	19	2497-4263
6	559-1076	20	2799-4737
7	645-1205	21	3100-5254
8	775-1335	22	3445-5857
9	861-1464	23	3832-6503
10	990-1636	24	4263-7235
11	1076-1808	25	4737-8053
12	1205-2024	26	5254-8957
13	1335-2239	27	5857-9948
14	1464-2497		

1, ..., 28). Each set of four filter bank outputs was transformed to 4 MFCCs and feature warping (as described in Section 2.2.5.3) [2] was applied.

Using the output of 4 adjacent filters for experiment on each sub-band was initially suggested by Besacier et al, and we used the same configuration to be able to compare our findings with theirs.

5.4.2 GMM-UBM SYSTEM

The SID and AID systems are built on the GMM–UBM method [80], please refer to Section 2.3.3 for more details. In what follows, "class" refers to accent or speaker, depending on the particular experiment.

In the GMM-UBM approach, using utterances from the training sets of all classes, a UBM is built. Class-dependent models are obtained by MAP adaptation (full description of this method is accessible from Section 2.3.3.2) [80], adapting the means of the UBM, using the class-specific enrollment data. The result is one UBM and C class-dependent GMMs, where in our experiments C is 14 and 93 or 94 (depending on the "jack-knife" set) for AID and SID, respectively.

Using the technique described in [95] (Please refer to Section 2.4.3 for more details) , for AID, the inter-session variability within a class, such as inter-channel and inter-speaker variability, is estimated. On a band specific level with a fixed band-independent rank this technique is applied. ISV modelling was used in the final AID sub-band experiments (Fig. 5.5.2 and Fig. 5.5.3) but not used in the full bandwidth AID experiments (Table 5.5.1) or the initial sub band experiments (Fig. 5.5.4 and Fig. 5.5.5).

During this research the clean microphone recorded speech files were used. Application of ISV modelling is investigated on AID as it will remove the unwanted speaker differences in each accent class, however for SID, speaker differences are what system is looking after. ISV modelling usually applies on SID using the telephony recorded speech, where the recording environments and transmission channels are different for recordings. So as in this research the clean speech are used ISV modelling was not used in the SID systems.

5.4.3 I-VECTOR SYSTEM

Motivated by the success of i-vectors in the field of speaker recognition, this chapter also proposes a approach for AID from telephone speech patterns based on i-vectors. In this method, each utterance is modelled by its corresponding i-vector. Then, SVM is applied to identify the accent of speakers.

The overall block diagram of the i-vector based accent identification system is illustrated in Figure 5.4.1. In this diagram x_N and y_N are representing extracted features and their corresponding labels, respectively. Blue and red parts of the diagram are corresponds to train and test phase of the accent identification system, respectively. The details about the i-vector system and SVM classifier are available in Section 2.4.2, and 2.4.1, respectively. The i-vector system which is used

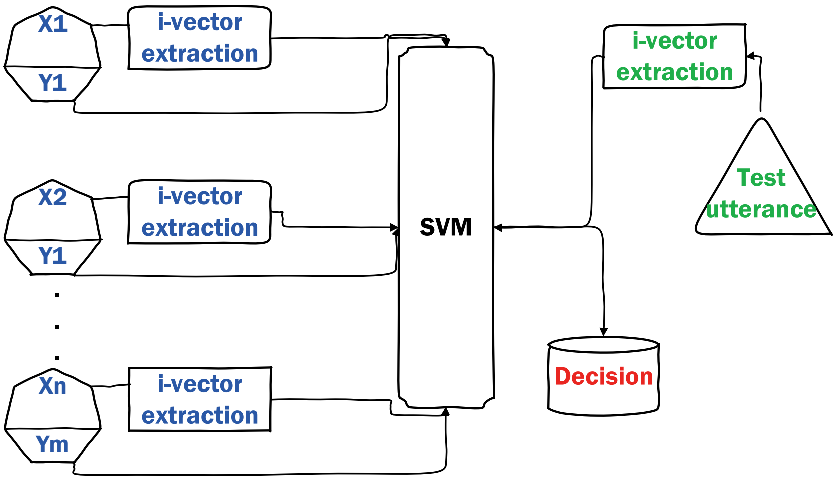


Figure 5.4.1: The block diagram of the Accent ID system based on the i-vector approach, depicting both training and testing phase. X_n and Y_n represent samples and class labels, respectively.

is identical as the one that is used for SR (Section 4.4), except that for AID the SVM is used for decision making.

This method is trained and tested on clean speech recordings from the same ABI database. Evaluation results show that the proposed method outperforms different conventional methods [105] for the problem of the automatic identification of speaker's accent from his/her speech sample.

Table 5.5.1: Summary of Results for SID and AID Systems (identification rate)

GMM	AID (30 sec)		SID (30 sec)		SID (10 sec)		SID (3 sec)	
comp.	Full	Tel	Full	Tel	Full	Tel	Full	Tel
512	38.50	57.50	100	97.54	100	95.09	98.98	88.18
2048	40.64	59.42						
4096	42.54	60.34						

5.5 Experimental Results and Discussion

In order to demonstrate the competitiveness of our AID and SID recognition systems, the first experiments are performed using the full bandwidth (0–11.025 kHz) and telephone bandwidth (0.23–3.4 kHz) speech. Table 5.5.1 presents the obtained results. The AID system here uses a pitch-based SAD. Using the full bandwidth speech, the performance of the AID system using 30 second test segments is 38.50%, 40.64% and 42.54% with 512, 2048 and 4096 mixture components, respectively. Using a 512 component GMM, the performance of the SID system is 98.98%, 100% and 100% for 3, 10 and 30 second test files, respectively. The AID performance increased by between 42% and 49% when using the simulated telephone bandwidth speech, whereas the SID performance dropped by between 3% and 11%. Current advanced GMM-UBM based AID systems usually use 4096 component GMMs and pitch-based SAD, which also achieved the best performance (60.34%) in our experiments (the corresponding performance for energy-based SAD is 57.37%).

In the following, the performance of the SID and AID systems for each individual sub-band will be investigated. For the purpose of analysis, dividing the entire frequency range into four broader regions is also useful: A from 0 to 0.77 kHz, B from 0.34 to 3.44 kHz, C from 2.23 to 5.25 kHz and D from 3.40 to 11.02 kHz. The energy-based SAD is used for both the AID and SID systems (as the pitch-based detector would eliminate most of the high-frequency unvoiced fricative sounds). As the feature dimensionality is much lower when using individual sub-bands as opposed to the full bandwidth and based on the results in Table 5.5.1, both the SID and AID systems in the following experiments are based on 512 component GMMs.

Figure 5.5.1 shows the SID performance as a function of frequency sub-bands. We can see that when using the mid frequency sub-bands (region B), the lowest performance is obtained. These results are compatible with earlier findings reported in [31], which were obtained for

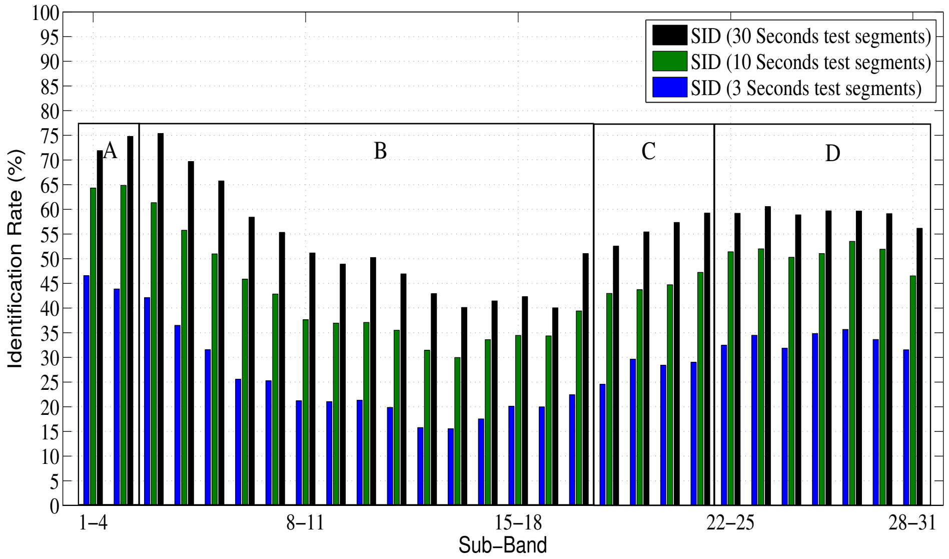


Figure 5.5.1: SID performance as a function of frequency sub-band for 3, 10 and 30 second test signals when using 512 component GMMs. Obtained performances when using full bandwidth, 512 mixture components GMM, and 3, 10 and 30 seconds test utterances are 98%, 100% and 100%, respectively.

clean speech on the TIMIT corpus and using only mono Gaussian modelling. The performances for 3, 10 and 30 second test files show similar trends, but accuracy is around 27% and 10% lower on average for the 3 and 10 second test data, respectively, compared to the 30 second data. Figure 5.5.2 shows the results for AID. Compared with the results for SID, region B seems to be more useful, while regions C and D are less useful. Dividing frequency regions into four was based on the results presented in Figure 5.5.3 and Figure 5.5.5.

In order to contrast the SID and AID performances, the results presented in Figure 5.5.1 and Figure 5.5.2 were first normalized to sum to one over all the sub-bands and then subtracted. Figure 5.5.3 shows the resulting contrastive SID and AID performance, which we refer to as normalised SID (NSID) and normalised AID (NAID), for 30 second test data. In Figure 5.5.3 Regions with positive values (A, C and D) contain more speaker specific information than accent information, whereas the region with negative values (B) carries more AID information. For SID, region A, corresponding to the primary vocal tract resonance information of vowel and nasal sounds, and regions C and D, corresponding to high frequency sounds such as frica-

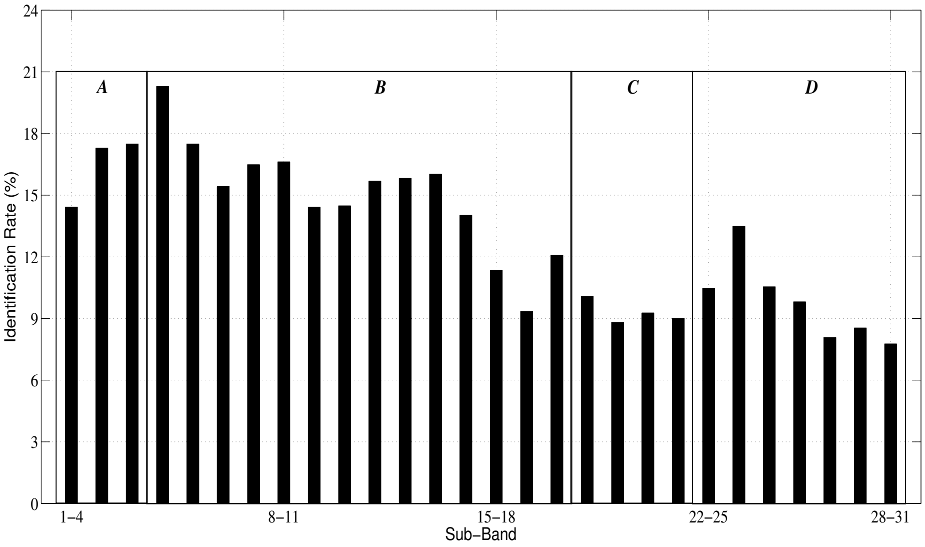


Figure 5.5.2: AID performance as a function of frequency sub-bands using 512 component GMMs and 30 second test signals. The obtained performance when using full bandwidth and 2048 mixture component is 40%.

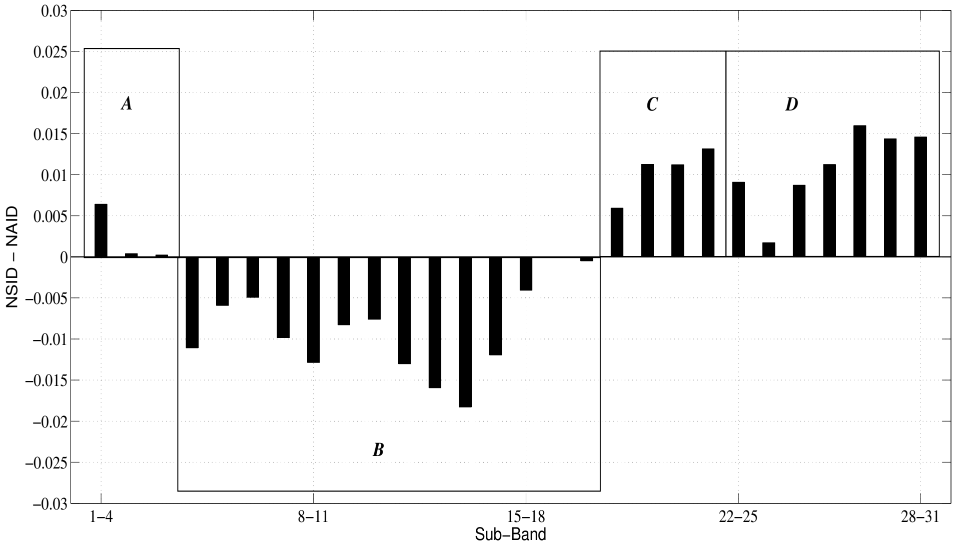


Figure 5.5.3: The difference between the normalized SID and AID performance for frequency sub-bands using 30 second test signals.

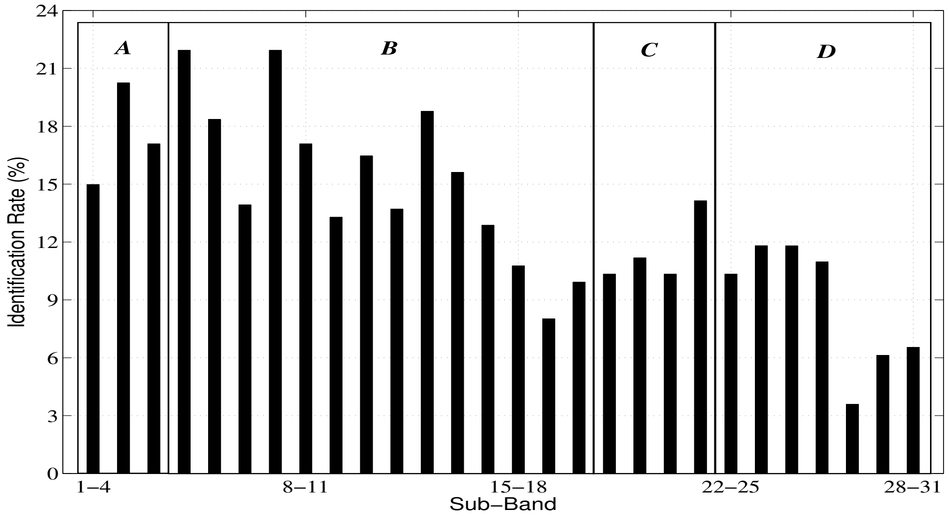


Figure 5.5.4: AID performance after ISV compensation as a function of frequency sub-bands using 512 component GMMs and 30 second test signals.

tives, are most useful. Where one would expect to find vocal tract resonance information for general voiced speech sounds is region B. While this information will be influenced by individual differences in vocal tract physiology, linguistic information dominates and makes the region most useful for AID. This is compatible with the observations on the importance of vowels in subjective analyses of accent [162].

Next the results obtained with ISV compensation [95] are presented. In order to inquire which frequency bands achieve most from ISV compensation in AID, we apply it to each frequency sub-band separately, with ISV compensation subspace dimension of 100. Figure 5.5.4 shows the result. Figure 5.5.5 shows normalized AID (after ISV compensation) subtracted from normalized SID. Comparing Figure 5.5.3 and Figure 5.5.5 indicates that ISVC gives the biggest gain in region C. In fact, comparing Figure 5.5.3 and Figure 5.5.5, the average improvement of AID performance in this region is 24%, compared with average improvements of 6% and 2% in regions A and B, respectively, and an 12% decrease in region D. This suggests that ISV compensation is capable of compensating for some of the speaker-dependent information in region C, which is noise from the perspective of AID, but not in region D.

Last, to use our findings form experiments on isolated sub-bands, and in order to improve

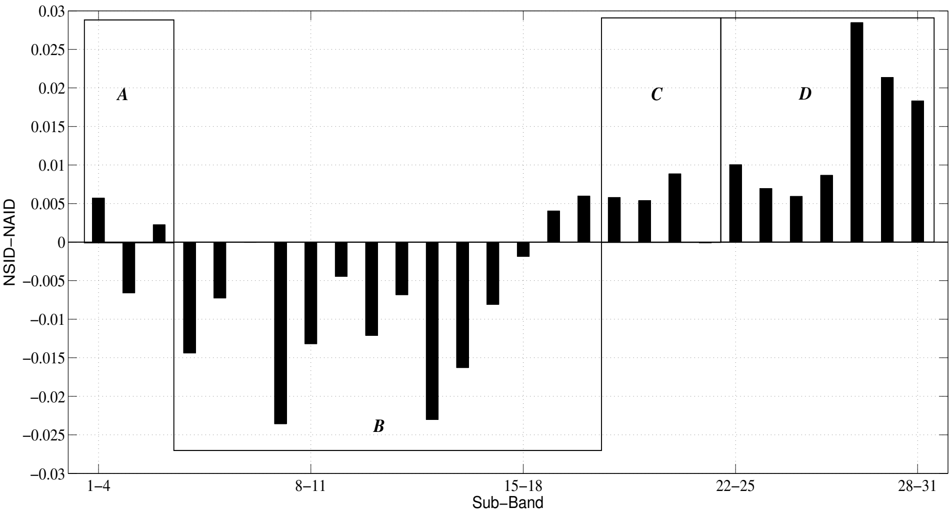


Figure 5.5.5: The difference between the normalized SID and AID performance for frequency sub-bands (after application of ISV compensation to AID system).

the performance of the automatic AID systems, the same training and testing material is used but instead of full-bandwidth, band-limited speech recording to 5.25*kHz* was used. This selection is based on the illustrated results in Figure 5.5.5 and 5.25*kHz* is corresponds to the centre frequency of the 25^{th} filter, which means regions A, B, and C are used in the new set of experiments. In these set of experiments the i-vector system is used for automatic identification of speaker's accent from his/her speech sample. Because of the similarities between ISV modelling and i-vector approach to find these regions Figure 5.5.5 is used instead of the Figure 5.5.3. Table 5.5.2 shows the average results for experiments using three folds of jack knife set and different configurations.

As shown in the Table 5.5.2 the best performance is achieved by using 512 mixture components, T-matrix of size 800 and speech which contains frequency up to 5.25 kHz, which is 76.76. To be able to compare the performances of i-vector and GMM-UBM system, the best configuration is used, but instead of band-passed filtered speech signals, we used the full bandwidth speech and the performance was 72.53%, which is still far better than the comparable performance achieved by the GMM-UBM system (42.54%).

Table 5.5.3 shows the confusion matrix for the experiment of Table 5.5.2, which is printed in bold text. From this confusion matrix the lowest automatic identification rate corresponds

Table 5.5.2: Summary of Results AID task using i-vector System and Band Limited speech up to 5.25 kHz and the best published results using single system and full-bandwidth speech

Proposed system	i-vector	i-vector	i-vector
	(256)	(512)	(1024)
Rank of T-matirx	ID (%)	ID (%)	ID (%)
200	-	68.00	70.80
400	-	74.30	74.30
800	-	**76.76**	75.30
Published result [108]			
300	68%	-	-

Table 5.5.3: Confusion matrix for the best i-vector system

	brm	crn	ean	eyk	gla	ilo	lan	lvp	ncl	nwa	roi	shl	sse	uls
brm	16	1	1	0	0	0	0	0	0	1	0	0	1	0
crn	0	11	1	1	0	3	0	0	1	1	0	1	1	0
ean	1	0	16	1	0	1	0	0	0	0	0	0	0	0
eyk	1	0	0	21	0	0	2	0	1	0	0	0	0	0
gla	0	0	0	0	19	0	0	0	1	0	0	0	0	0
ilo	2	0	2	1	1	12	3	0	0	0	0	0	0	0
lan	1	0	1	0	0	1	16	0	1	1	0	0	0	0
lvp	0	0	0	0	0	0	1	17	0	2	0	0	0	0
ncl	0	0	1	0	1	0	2	1	13	0	0	2	0	0
nwa	1	2	0	4	0	0	1	0	1	11	0	1	0	0
roi	0	0	0	0	0	0	0	0	0	0	15	0	0	4
shl	0	0	0	0	0	0	0	0	0	0	0	22	0	0
sse	0	0	0	2	0	2	0	0	0	1	0	0	11	0
uls	0	0	0	0	0	0	0	0	0	0	2	0	0	18

to *nwa*, which is 71%. The best identification is achieved for both *lan* and *shl* by only one miss identification.

From Table 5.5.3 it is obvious that the two accents which seems to be different from others and relativity similar to each other are *roi* and *uls*, with only 4 and 2 misidentified tests from 19 and 20 test utterances, respectively. The reason for this claim is that the miss identified test utterances from *roi* and *uls* accents are identified as having *uls* and *roi* accents, and none of them identified as having the accent from the other available regional accents.

5.6 Summary

By using the Accents of the British Isles speech corpus, this chapter investigated the effect of different frequency bands on automatic AID and SID. Both the AID and SID systems were based on GMM-UBM approach. The competitiveness of our systems was demonstrated by applying full bandwidth (0–11.025 kHz) and band-pass filtered (0.23–3.4 kHz and 0.23-5.5kHz) speech experiments.

The beginning of this chapter reports the results of applying GMM-based classifiers to AID (14 classes) and SID (93 or 94 classes) on the ABI speech corpus [145]. Using full bandwidth (11.025 kHz) speech and 512 component GMMs, SID accuracy is close to 100%, confirming that SID is a fairly simple task for this type of data [31].

The best AID accuracy, using GMM-UBM system, was 60.34%, which was obtained by using the band-pass filtered data, pitch-based SAD and 4096 component GMMs.

The experimental results contrasting the utility of information in narrow sub-bands for the AID and SID tasks revealed that dividing the spectrum into four regions is useful: A (0 to 0.77 kHz), B (0.34 to 3.44 kHz), C (2.23 to 5.25 kHz) and D (3.40 to 11.02 kHz). Our investigations verified that speaker information dominates in regions A, corresponding to primary vocal tract resonance information, and D, corresponding to high-frequency sounds. In contrast, as the vocal tract resonance information in region B is biased towards linguistic, rather than speaker information, region B is most useful for AID. While speaker information appears to dominate, region C contains both types of information. The biggest gain is observed in region C when ISV compensation is applied to the AID system, where AID performance is improved by 24%, indicating that ISV compensation is able to factor out some of the speaker information in this region.

Finally we used the band limited speech recordings of up to 5.25kHz for AID using an i-vector approach. The results, 76.76 % accuracy. In order to compare the performances the i-vector experiment is repeated with the same configurations except that in the new experiment full bandwidth speech is used as an input. The results confirm that that the i-vector system with the ID rate of 72.53 % outperform our GMM-UBM system (with compatible configuration) with the ID rate of 42.54%.

"The knowledge of anything, since all things have causes, is not acquired or complete unless it is known by its causes."

Avicenna (Ibn Sīnā)

6 Speaker Recognition for Children's Speech

Although automatic recognition of children's speech has been the subject of considerable research effort, there is little published work on issues and algorithms related to automatic recognition of a child's identity from their speech. For example, before this research [37] we did not know how increases in inter- and intra-speaker variability for children's speech [163] would affect SR performance or the significance of different frequency bands of SR for children, although the relevant studies for adult SR have been reported [31].

6.1 Introduction

The applications cover almost all the areas in which it is desirable to secure actions, transactions, or any type of interactions by identifying or authenticating the person making the transaction: for example, in secure facility access control [164], securing the access to reserved services (Telecom network, databases, and websites.), authenticating the user making a particular transaction (e-trade or banking transaction) and games [165]. Usually, a person claims their identity using different input methods, such as using an ID card or entering an ID code onto a key pad, and they will be asked to utter a specific phrase. Then the input speech sample will be compared to a reference model corresponding to the specific claimant, and based on the result, the decision will be made to either accept or reject the access.

6.1.1 Related works

To the date of this research, the success of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) and GMM-Support Vector Machine (GMM-SVM) approaches to adult SR motivated us to apply these techniques to our child SR task. More recently, as the application of factor analysis to adult SR offered a better recognition rate, the application of this new technique to child and adult SR tasks are also investigated.

The distribution of acoustic feature vectors for a population of speakers is typically captured using a UBM (a speaker-independent GMM constructed using data from a variety of speakers and background conditions) [80, 152]. Speaker-dependent GMMs are then built by MAP adaptation of the UBM [153]. At the recognition stage, each test segment is scored against speaker models, and the description of this system is included in Section 2.3.3.2. However, discriminative approaches, such as SVMs, can be used, which have been shown to obtain comparable, and in some cases better, performance than GMM-based systems. The combination of GMM super-vectors, comprising the stacked parameters of the GMM components with SVMs has also been successful (please refer to Section 2.4.1.2 for system description) [83]. Alternatively, research by Dehak et.al [87], which is described in Section 2.4.2, proves that the use of low dimensional identity vectors (i-vectors) as new feature sets, will provide better recognition rates, even using a simple scoring method. Score normalization approaches are usually applied for speaker verification tasks and they have been proven to be capable of boosting the verification performance by normalizing score distribution (please refer to Section 2.6 for more details).

6.1.2 Scope of the chapter

This chapter presents the experiment's results in SR for both children and adults, and is organized as follows: Section 6.2 describes the data. Our SR systems are described in section 6.3, and our experiments and results are presented in section 6.4. In the same section, we also describe a study of the information utility in different frequency bands for both children and adult SR and the problem of identifying a child in school and classrooms (by classroom we mean defining different number of classes i.e. speakers in train and test sets).

6.2 Data Description

Three corpora of British English speech and a corpus of American speech were used in this research: The PF-STAR [4] (British English speech) and OGI [141] (American English speech) corpora of children's speech, and the two ABI corpora of regionally-accented adults speech: ABI-1 [145], ABI-2.

6.2.1 PF-STAR kids' speech corpus

A full description of the PF-STAR kids' speech corpus is given in Section 3.2.2. From the entire corpus, all data from 150 speakers (4-14 years old) were used (for both training and testing); the remaining 8 speakers were the youngest children and did not record sufficient data to be included in the experiment.

A 10-second speech segment (after removing silence) from an utterance for each speaker is extracted for testing, and the rest of the utterances from speakers are used for training.

6.2.2 OGI kids' speech corpus

The OGI kids' speech corpus is fully described in Section 3.2.1. Four different test sets (10 seconds per utterance) from the OGI data are used in the experiments presented in this chapter.

TS1: To investigate the effect of different frequency bands on SR performance for general children's speech, 359 speakers were chosen randomly (from kindergarten to 10th grade).

TS2: To investigate the effect of different frequency bands on SR performance for speech from children of different ages, 3 different age groups were selected, each containing 288 speakers. These are AG1: kindergarten to 2nd grade (5-8 years old), AG2: 3rd to 6th grade (8-12 years old), and AG3: 7th to 10th grade (12-16 years old).

TS3: To investigate the problem of identifying a single child in a school, two 'schools' of 288 randomly-chosen speakers from kindergarten to 10th grade were chosen.

TS4: To investigate the problem of identifying a single child in a classroom (schoolroom), 12 'classrooms' of children from 3 grade groups were chosen, each containing 30 children (each classroom consisted of children from the same age-group).

6.2.3 ABI adult speech corpora

Full descriptions of the ABI speech corpora are given in Section 3.3.3. 4 hours of speech data from ABI-1 corpus is used to train the UBM to investigate the effects of bandwidth on a child's performance (from PF-Star data set) and an adult SR task. In order to provide a baseline against which the children's SR performance could be compared, the test speakers were taken from the ABI-2 corpus for adults and the PF-STAR corpus for children. From the ABI-2 data-set, 10-second segments of the test data were selected from each of 150 randomly-chosen target speakers.

6.3 Speaker Recognition Systems

6.3.1 Signal Analysis

Feature extraction was performed as follows: Periods of silence were discarded using an energy-based SAD. Speech was then segmented into 20-ms frames (10-ms overlap), and a Hamming window was applied. After applying the Hamming window, the short-time magnitude spectrum, which is obtained by applying an FFT, is passed to a bank of 24 (for f_s=16kHz) and 32 (for f_s=22.05kHz) Mel-spaced triangular bandpass filters, spanning the frequency region from 0Hz to $f_s/2$ Hz. f_s is the sampling frequency, which is 16kHz for OGI kids corpus and 22.05kHz, for ABI and PF-STAR. Table 6.3.1 and 5.4.1 shows the centre frequency of each filter (the cut-off frequencies of a filter are the centre frequencies of the adjacent filters), for both 16kHz and 22.05kHz sampling frequencies, receptively.

To investigate the effect of different frequency regions on SR performance, experiments were conducted using two methodologies. The first method is used during experiments on the PF-STAR and ABI corpora, and the second method is applied to the OGI corpus.

The first method is used to study the effect of bandwidth on verification performance for adults and children. To achieve bandwidth reduction, a 31 band-pass filter-bank analysis was performed, but the vector passed to the Discrete Cosine Transform (DCT) for calculation of the cepstral features (consisting of 19 static MFCCs and 19 delta MFCCs) consisting of different numbers of logarithm filter-bank energies, varying from 21, corresponding to the bandwidth of 3.6kHz, to 31, corresponding to the maximum bandwidth of 11.025kHz.

For the full bandwidth experiments, each speech frame was then represented as a 38 dimen-

Table 6.3.1: The Centre Frequencies for 24 Mel-spaced Band-Pass Filters

FILTER NUMBER	CENTER FREQ. (Hz)	FILTER NUMBER	CENTER FREQ. (Hz)
1	156	13	1843
2	281	14	2062
3	406	15	2343
4	500	16	2656
5	625	17	3000
6	750	18	3375
7	875	19	3812
8	1000	20	4312
9	1125	21	4906
10	1281	22	5531
11	1437	23	6281
12	1625	24	7093

sional feature vector, consisting of 19 static MFCCs and 19 delta MFCCs.

In the second method (which is very similar to the method which is used in Section 5.4.1), to investigate the effect of different frequency regions on SR performance, experiments were conducted using frequency band limited speech data comprising the outputs of groups of only 4 adjacent filters, using the OGI kids speech corpus. We considered 21 overlapping sub-bands, in which the N^{th} sub-band comprises the outputs of filters N to N+3 (N=1 - 21). Each set of 4 filter outputs was transformed to 4 Mel Frequency cepstral coefficients (MFCCs) and mean and variance normalization [2] were applied. For the full bandwidth experiments (using the OGI corpus), the outputs of all 24 filters were transformed into 19 MFCCs. In order to investigate the effects from inclusion of delta coefficients, another experiment was designed with exactly the same configuration, for full band experiment. The only difference is that the delta coefficients are also added to the feature vectors. So by including deltas, each speech frame is represented by 38 dimensional feature vectors, as against with 19, which is used initially.

Feature Warping [2] with a 3-second window is applied to the MFCC feature vectors to reduce the effect of channel mismatch and additive noise for the first and second methods.

6.3.2 Modelling

Our SR systems are based on the GMM-UBM [83, 153], GMM-SVM [83] and factor analysis methods [87], which are described in Section 2.3.3.2, Section 2.4.1.2, and Section 2.4.2, respectively.

6.3.2.1 Modelling approaches for experiments using the OGI kids' corpus

When using the OGI kids corpus in the GMM-UBM approach, a UBM is built using utterances from all data in all speakers' training sets. Speaker-dependent models are obtained by MAP adaptation (adapting means only) of the UBM, using 48-second segments of speaker-specific enrolment data. The result is one UBM and 1083 speaker-dependent GMMs (a small number of speakers for whom there was very little data were not used).

In our GMM-SVM system, the training data from each individual speaker was divided into three segments (each 16 seconds in length) and each was used to estimate the parameters of a GMM by MAP adaptation of the UBM (using the relevance factor 10 [100], please refer to Section 2.3.3.2 for more details). The adapted GMM mean vectors are then concatenated into a super-vector [83] (full details about this technique is included in Section 2.4.1.2), and the speaker classes are assumed to be linearly separable in the supervector space. The super-vectors are used to build one SVM for each speaker by treating that speaker as the 'target' class and the others as the 'background' class.

In the i-vector approach, as in the GMM-SVM approach, three segments of speaker-dependent speech are used to MAP adapt the UBM means in order to obtain the speaker-dependent super-vectors. The same speech utterances which are used to train the UBM are also used to train the total variability subspace, as described in Section 2.4.2.2. In the factor analysis model, each of these data points in the super-vector space are imagined to be generated by sampling a k-dimensional multivariate Gaussian ($k < super - vectordimension$). So instead of using super-vectors as features, we used the mean vector (i-vectors) of trained k-dimensional multivariate Gaussian distribution. These low dimensional mean vectors are believed to control the principal dimensions of the total variability space.

In our GMM-UBM, GMM-SVM, and i-vector automatic recognition systems, the score for each speaker model is normalized using the highest score across all speakers (max-log-likelihood score normalization), using the Equation 2.48 in Section 2.6.

6.3.2.2 Modelling approaches during experiments using the ABI and PF-STAR corpora

The GMM-UBM and GMM-SVM systems are used for experiments on the PF-STAR and ABI speech corpora. Descriptions of these systems are included in Section 2.3.3.2, and Section 2.4.1.2, respectively.

Two gender-independent UBMs were trained using approximately 4 hours of speech data from each of the ABI-1 and the PF-Star corpora with 10 iterations of the EM algorithm for adult and children, respectively.

In total, 152 adult speaker-dependent GMMs and 150 children speaker-dependent GMMs were trained. The speaker-dependent GMMs for adults and children were obtained by applying MAP-adaptation to the means of the GMM-UBM using the relevance factor 10 [100] (please refer to Section 2.3.3.2 for more details). In all cases, the adaptation was performed using approximately 48 seconds of speech data from each subject (after silence removal).

In our GMM-UBM and GMM-SVM recognition systems, the score for each speaker model is normalized using the highest score across all speakers (max-log-likelihood score normalization), using the Equation 2.48 in Section 2.6.

6.4 Experimental Results and Discussion

6.4.1 Verification and identification experiments

Verification experiments were conducted using a version of the methodology developed for the NIST speaker recognition evaluations. Each test utterance was scored against the ‘true’ (correct) speaker model and 10 ‘impostor’ models. Results are presented in terms of percentage EER, calculated using the standard NIST software.

Identification experiments involved scoring each test utterance against a fixed test set (or all the speaker models) of speaker models and assigning the test utterances to the class with the highest score.

6.4.2 Experimental results of PF-Star and ABI corpora

6.4.2.1 Comparison of speaker verification performance for adult and child speech

In this section, we first analyse speaker verification performance for adults and children for various numbers of mixture components, using the maximum 11.025kHz bandwidth. We will also study the effect of bandwidth on verification performance for both adult and children speakers.

6.4.2.2 The effect of the number of mixture components

The results of experiments on the effect of the number of mixture components are presented in Figure 6.4.1. It can be observed that the best performance for adults is 0.22% EER with 128 or 256 mixture components. This observation confirms that speaker verification for adults using clean, wide-band speech is a relatively easy task [166]. For children the best performance is 0.8% EER, also obtained with 128 and 256 mixture components. This indicates that the speaker verification EER for children is nearly four times worse than for adults.

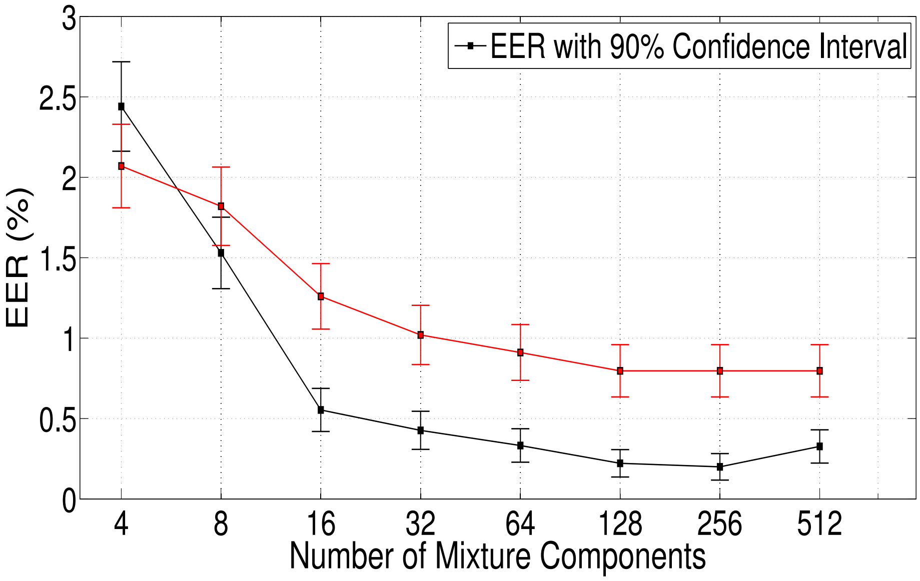


Figure 6.4.1: Speaker verification performance in terms of equal error rate (EER) for adult speech (in black) and child speech (in red) when using the full bandwidth and various numbers of mixture components and using ABI and PF-Star corpora.

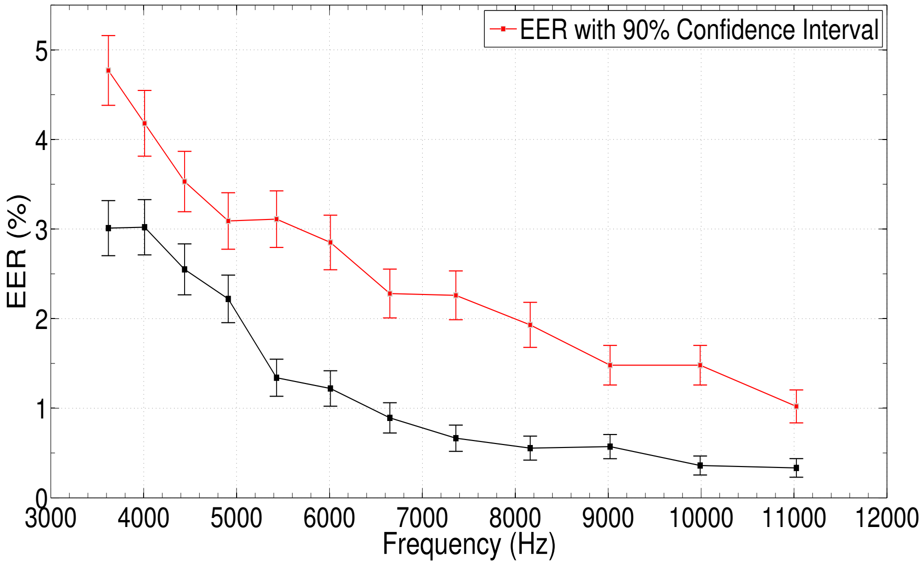


Figure 6.4.2: The speaker verification performance in terms of equal error rate (EER) for adult speech (in black) and child speech (in red) as a function of the bandwidth of speech signal.

6.4.2.3 Effect of bandwidth (first method)

In this section, we study the effect of bandwidth on verification performance for adults and children. From the previous section, it is clear that the EER for the adult data in our study is low when 128 mixture components are used. Therefore, in order to obtain results that are statistically more reliable, the experiments in this section are performed for both adult and children, using GMMs with just 32 mixture components. This is consistent with [166, 167], in which 32 component GMMs were also used and performed well on TIMIT.

To achieve bandwidth reduction, the same 32 band-pass filter-bank analysis from the previous experiments was performed, but the vector passed to the DCT for calculation of the cepstral features consisted of different numbers of logarithm filter-bank energies, varying from 21(corresponding to the bandwidth of 3.6kHz) to 32 (corresponding to the maximum bandwidth of 11.025kHz).

Verification results in terms of EER for adults and children as a function of the bandwidth are shown in Figure 6.4.2. For adults (Figure 6.4.2, red graph), it is evident that it is useful to partition the spectrum into three regions: (i) up to 3.5-4kHz, (ii) 3.6-4kHz to 5.5kHz, and (iii)

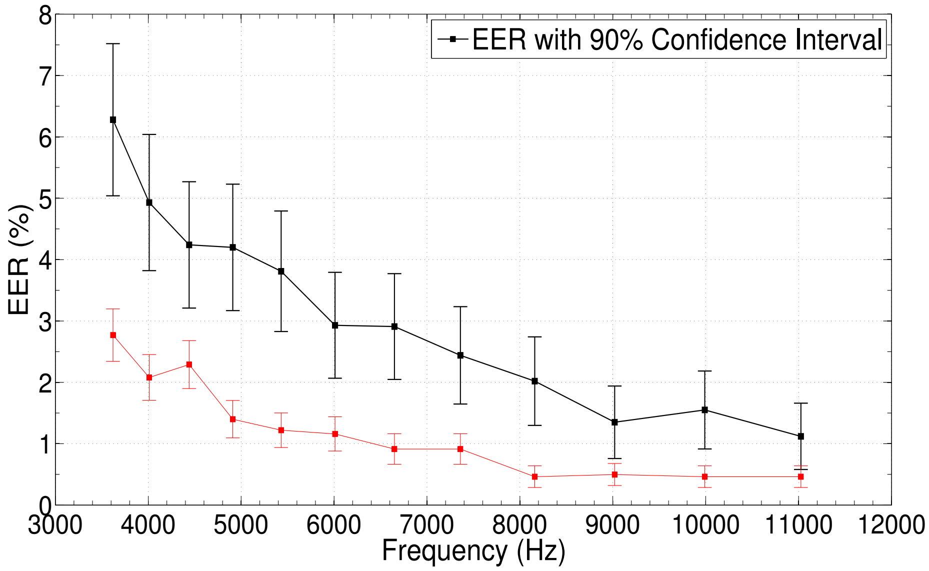


Figure 6.4.3: Speaker verification performance in terms of equal error rate (EER) for younger children aged from 5 to 9 years (in black) and older children aged from 10 to 14 years (in red) as a function of the bandwidth of speech signal.

above 5.5kHz. Region (i), corresponding to the vocal tract's primary resonances, clearly contains speaker-specific information. However, in these experiments for adults, there appears to be no benefit from including frequencies above 3.6kHz in this region. Region (ii) contributes a 58% reduction in EER. The importance of this region for speaker verification has been noted in [167]. Region (iii) accounts for a further 76% reduction in error rate but over a much larger frequency range. The importance of this region for speaker recognition has been noted in [36]. For children there are clearly two slopes to the graph above 3.6 kHz, increasing the bandwidth between 3.6 and 5.5 kHz gives a gain of about 1% EER for each kHz of bandwidth added, whereas the gain between 5.5 kHz and 11 kHz is lower, about 0.2% for each kHz added.

Figure 6.4.2 (red graph) shows the corresponding results for children's speech. However, a clearer picture emerges from figures 6.4.3 (black graph) and 6.4.3 (red graph), where the results for younger children (aged 5 to 9 years) and older children (aged 10 to 14 years) are presented separately. For younger children, the boundary for region (i) appears to be between 4.5 and 5.5kHz. In contrast with the case for adults, there is useful information in the 3.6 to 4.5kHz region, presumably because the primary vocal tract resonances occur at higher frequencies for children with smaller vocal tracts. Region (ii) lies between 4.5-5.5kHz and 6.5kHz, approximately 1kHz higher than for adult speech, and contributes a 37% reduction in EER. It would be interesting to discover if this is consistent – in terms of physiology – with the corresponding result for adult speech. Region (iii), comprising frequencies above 6.5kHz, contributes a 64% reduction in EER over 4.5kHz. The results for older children (figure 6.4.3(red graph)) are similar to those for adults.

6.4.3 Experimental results on OGI kids speech corpus

6.4.3.1 Full-bandwidth SR for children's speech

Table 6.4.1 shows the results of SR experiments on full-bandwidth speech for the 3 age groups of children (AG1 to AG3, 288 children per group), using 1024 component GMM-UBM and GMM-SVM systems and a 64-component GMM-SVM system, with and without delta-features. These sizes of GMM were found empirically from the evaluative experiments. The last column of this table contains the result of McNemar's test for identification experiments for different age-groups.

Table 6.4.1: SR performance for three different grade groups (AG1, AG2 and AG3).

	GMM-UBM	GMM-SVM	GMM-SVM	GMM-SVM	Statis tests
	(1024)(MFCC)	(1024)(MFCC)	(64)(MFCC)	(64)(MFCC + Δ)	McNemar's
Verif.	EER (%)	EER (%)	EER (%)	EER (%)	p-value
AG1	02.10	06.94	02.00	**01.80**	-
AG2	01.33	03.48	01.04	**00.21**	-
AG3	00.67	02.83	00.84	**00.64**	-
Identif.	ID (%)	ID (%)	ID (%)	ID (%)	p-value
AG1	62.15	38.54	75.00	**75.00**	AG1AG2
					0.009012
AG2	80.56	79.17	88.19	**89.24**	AG2AG3
					0.1379
AG3	85.71	83.33	93.06	**93.26**	AG1AG3
					0.00118

Both identification rate and EER improve as the children's ages increase. For example, the EER falls by 70% from 2.1% for the youngest to 0.64% for the oldest children. The corresponding increase in identification rate is 38%. The performance of the 1024 component GMM-SVM system was unexpectedly poor. An experiment on a separate evaluation set showed that the best number of GMM components for this system is 64, due to the short test utterances. The performance of the 64-component GMM-SVM system is shown in column 4,5 of Table 6.4.1. Verification performance is similar to that obtained for the 1024 component GMM-UBM system, but the identification rates are between 9% and 20% better for the 64 component GMM-SVM system. Based on the results of statistical significance test, presented in the Table 6.4.1, the deferences between AG1 and AG2, and also between AG1 and AG3 are statistically significant, but differences between AG2 and AG3 seems to be relatively small and the performance differences could be due to the chance.

6.4.3.2 Experiments on isolated sub-bands (second method)

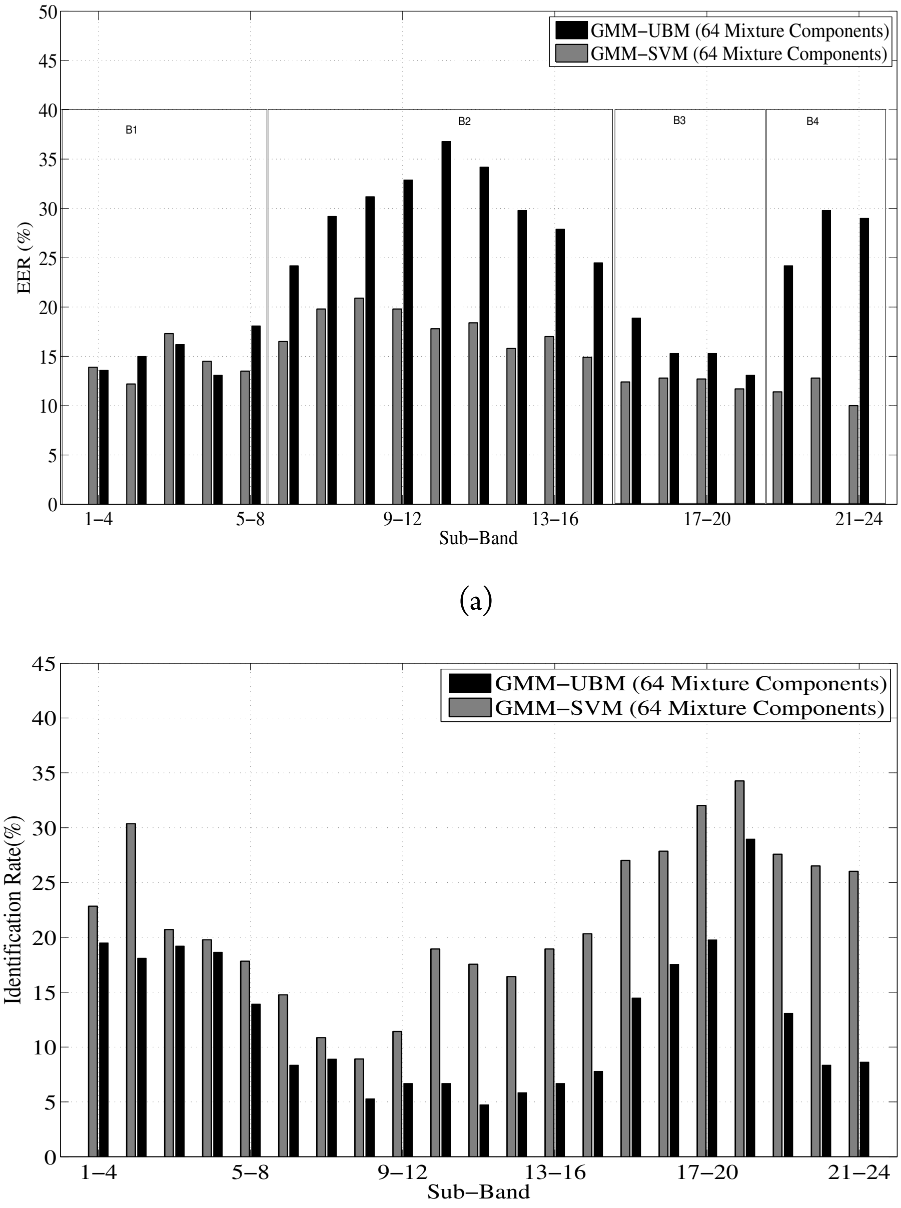
In this section, we study the effect of different sub-bands on verification and identification performance for children's speech from the OGI corpus. SR tests are conducted separately on 21 sub-bands, each consisting of four consecutive channels (please refer to Section 6.3.1 for more details).

Figures 6.4.4(a) and (b) show the verification and identification performances, respectively, for the 359 speaker test set (TS1) on each of the 21 sub-bands, using 64-component GMM-UBM and GMM-SVM systems (64-component GMMs were found to be adequate for these 4 dimensional sub-bands). Overall, it is clear that the GMM-SVM approach outperforms GMM-UBM. Based on our experiments we have found that by having enough training data and long enough test utterances GMM-SVM always performs better than GMM-UBM.

In the case of verification, Figure 6.4.4(a) shows sub-band EERs varying between 10% and 37%. For identification (Figure 6.4.4(b)) the sub-band identification rates vary between 5% and 34%.

Figure 6.4.5 shows the correlation among identification rates achieved using GMM-UBM and GMM-SVM systems. Histograms of the variables appear along the matrix diagonal; scatter plots of variable pairs appear off diagonal. The slopes of the least-squares reference lines in the scatter plots are equal to the displayed correlation coefficients. Obtained correlation coefficient for these two measure is +0.70 and it suggests that two measures tend to vary together.

From Figure 6.4.4, it is evident that, as in the case of adult speech [31], it is convenient to



(a)

(b)

Figure 6.4.4: Sub-band speaker verification rate (EER) (a), and speaker identification rate (b) for child speech from OGI corpus for different frequency bands.

partition the spectrum into 4 frequency regions, B1 to B4, where B1 corresponds to sub-bands 1-5 (0-1.13kHz), B2 to sub-bands 6-14 (0.63kHz to 3.8kHz), B3 to sub-bands 15-18 (2.1kHz

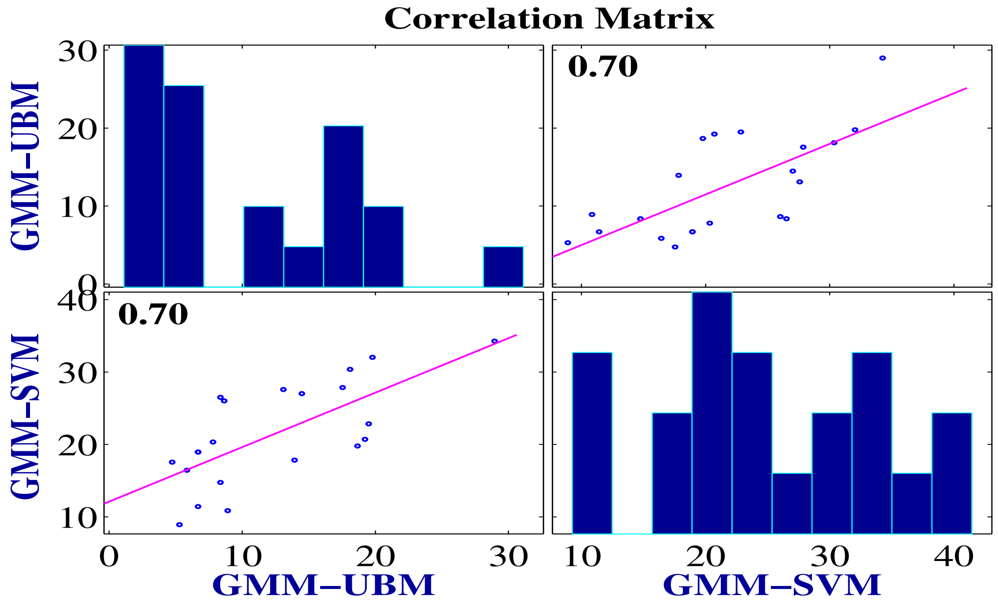


Figure 6.4.5: Scatter plots, least square reference line in the scatter plots, and histogram of the variables (identification rates from GMM-UBM and GMM-SVM systems).

to 5.53kHz), and B4 to sub-bands 19-21 (3.4kHz to 8kHz). The most useful bands for SR are B1, which contains individual differences in the part of the spectrum due to primary vocal tract resonances and nasal speech sounds, and B3, which contains information relating to high-frequency speech sounds such as fricatives.

Interestingly, the GMM-SVM system is able to extract more speaker-specific information from B2 than the GMM-UBM system. The importance of fricatives (hence region B3) for SR has been noted previously in [168]. Frequency regions similar to B1 to B4 were identified in [31] for adult SR on TIMIT. However, compared to the adult values, the frequency ranges spanned by these bands for children's speech are increased by approximately 38% (B1), 21% (B2) and 11% (B3). Comparison of the obtained results from first and second methods (Figure 6.4.3 (black) and Figure 6.4.4 (b)) of investigating the effect of frequency regions on the performance of speaker recognition using child speech, confirms the importance of high frequency regions for speaker recognition tasks using children speech.

Figure 6.4.6 shows sub-band speaker identification rates for 3 different age-groups of children, namely AG1, AG2, and AG3 (described in section 3.2.1). The figure shows that in almost all cases, the best performance is provided by the older children, and identification rate

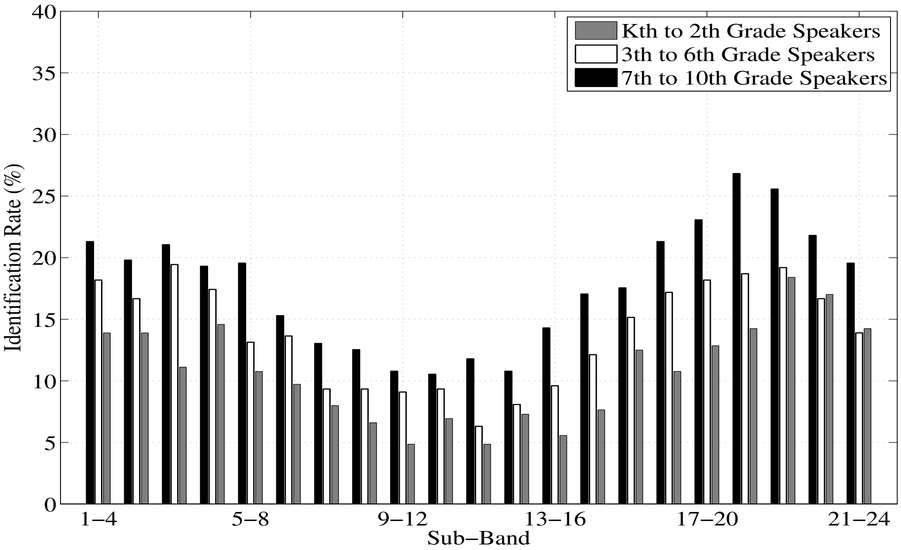


Figure 6.4.6: Sub-band speaker identification rates for three age groups of children, namely AG1, AG2 and AG3. The obtained performances using full-bandwidth speech signal was 75%, 89%, and 93%, for AG1, AG2, and AG3, respectively.

decreases for younger children. The figure shows the same decrease in performance between B1 and B2, and an increase between B2 and B3, for all 3 age groups. However, one would expect these changes to take place at higher frequencies for younger children because in general, younger children have shorter vocal tracts and smaller vocal folds (for young children with short vocal tracts, formants and other structures will occur at higher frequencies). Close inspection of figure 6.4.6 indicates that this is the case.

The results regarding the oldest children (7^{th} to 10^{th} grade, AG3) is consistent with published results for adult speaker identification on TIMIT [31].

6.4.3.3 Recognizing an individual child in a classroom and school

The purpose of this experiment is to evaluate SR performance for children's speech on tasks which are representative of potential applications. Table 6.4.2 shows the results of using different systems to recognize an individual child in a classroom (30 children from the same grade group as the target child) or school (288 children uniformly distributed across grades). The 'classroom' experiment is conducted for simulated classrooms from age groups AG1, AG2, and AG3. For each age group, the experiment was repeated for 4 random simulated classrooms, and the average result is given in Table 6.4.2.

Table 6.4.2: SR performance for three different grade groups (AG1, AG2 and AG3). SV-D stands for Supervector dimension and T-D is the dimension of the T-matrix.

	GMM-SVM(64)		I-Vector(256)	
	SV-D=3648		SV-D=14592,T-D=400	
SR Performance	EER (%)	ID (%)	EER (%)	ID (%)
Classroom(AG1)	01.92	89.99	-	-
Classroom(AG2)	01.04	95.83	-	-
Classroom(AG3)	**00.83**	**99.16**	-	-
School(K^{th}-10th)	01.74	81.00	**01.00**	**87.15**

The results show that a child in a classroom is identified with accuracies of approximately 90%, 96%, and 99% for classes of 30 children in age groups AG1, AG2, and AG3, respectively. The McNemar's test is performed on the performances of AG1 and AG3, and the result was, p-value = 0.0492. So as with speech recognition, speaker recognition appears to be more difficult for younger children.

The identification rates for an individual child in a school of 288 children, using GMM-SVM and factor analysis feature modelling, are 81% and 87.15%, respectively.

6.5 Summary

This chapter presents the results of experiments in SR for both children and adult speech. Because of relevantly small research on detection technologies for children speech, we could not find any comparable system. But based on the some of the research on ASR for children, which recently reviewed in Section 2.1.1 and Section 1.4, it was initially expected to get worse identification rate when using children speech, and it was also expected to see interesting trend for the performances obtained using different sub-bands.

In the first part of this chapter, we compared speaker verification performance for adults and children, and in both cases, investigated the effects of bandwidth on EER. There are some differences between databases which were used for adult and children experiments, which make the comparison hard, but the best EERs which obtained for children and adults are 0.8% and 0.22%, respectively (using two databases). This data could suggests that any advantage stemming from increased inter-speaker variability in children is countered by the increase in intra-speaker variability.

Turning to bandwidth, we found, as reported elsewhere, that in terms of its contribution to speaker verification performance, the spectrum can be usefully partitioned into three frequency bands. For adult speech these are: (i) up to 3.5-4kHz, (ii) 3.5-4kHz to 5.5kHz, and (iii) above 5.5kHz. Similar bands occur for child speech, but with boundaries that are approximately 1kHz greater than for adults. A study of the utility of different narrow frequency bands – using the second method – for child SR has shown that as with adults, the spectrum can be usefully partitioned into 4 regions referred to as B1, B2, B3, and B4. Most useful speaker information is concentrated in B1, which contains the primary vocal tract resonances, and B3, which contains high-frequency speech sounds such as fricatives. However, the frequencies at which these regions occur are between 11% and 38% higher for young children than for adults. It has also been shown that sub-band SR identification rates are consistently poorer for younger children than for older children.

Experiments which simulate recognition of an individual child in a classroom or a school containing 30 (in total 4 simulated classroom for each age-group) and 288 children, respectively, using a 64-component GMM-SVM system, show that identification rates for a child in a class vary between 90% for the youngest to 99% for the oldest children, and that the identification rate for a child in a school is 81%. The school performance improved when using the i-vector system by 6.15% from 81% to 87.15%.

"Youth has no age."

Pablo Picasso

7

Identification of Age-group and Gender from Children's Speech

7.1 Introduction

Research effort into paralinguistic speech processing has been growing considerably over the last two decades. It has initially focused mainly on speaker recognition from adults' speech, e.g., [169], but more recently has also spread to speaker recognition for children's speech, e.g., [37] (please refer to Chapter 6), recognition of accent, e.g. [36, 105, 159], emotions, age, and gender e.g. [118, 137, 140]. A recent review of paralinguistic speech processing is presented in [105]. Some earlier research on gender and age recognition using adult speech demonstrated high performance [170, 171].

Automatic recognition of paralinguistic information for children can be beneficial in many application areas. Some of these applications are presented in Chapter 6. It could also be employed to adapt speech models, to guide a child computer interaction system to automatically adapt content, to enhance child security and protection, or in a wide range of educational applications. For instance, some social networking sites are designed specifically for children, (for example, "Club Penguin" - please visit http://www.clubpenguin.com for more details). As such systems evolve to include speech, an automatic system that recognizes the age, gender, or identity of a person from their voice could be a valuable safeguard for a child engaged in social networking. For example, this can be effective in providing protection from an adult masquerading as a child. In education, an interactive educational tutor could recognise the age and gender of a child and adapt their content appropriately.

7.1.1 Related works

Many studies focus on exploring the use of features capturing different types of information from the speech signal and the use of different classification methods, as pointed out in Section 2.2, and most of these employed MFCCs. The use of temporal patterns (TRAPs) features to capture longer temporal context was explored in [118]. Several studies also considered the use of glottal and prosodic features. These were typically calculated on the whole utterance and included features such as the fundamental frequency, articulation rate, and harmonic-to-noise ratio [118, 137, 172]. The latter could also be provided by estimating the spectral voicing information using the method presented in [173]. Overall, the use of MFCC features, capturing vocal-tract information, was shown to provide the best performance, which could be further improved by incorporating other features or combining multiple classifiers. The use of various classification approaches for age and gender identification has been explored. Early studies employed distance measures [170] and GMM and HMM based recognisers [171]. More recently, the success of GMM-UBM (system description is included in Section 2.3.3) and GMM-SVM (system description is included in Section 2.4.1.2) approaches to adult speaker recognition motivated its application to the age and gender recognition tasks [115]. The GMM-UBM and GMM-SVM approaches have also been compared to the use of GMM and parallel phone recognition systems [115, 137, 174]. The use of dynamic Bayesian networks employing prosodic features was explored in [137, 174]. Furthermore, techniques such as cepstral mean subtraction and variance normalization have been applied to speaker gender identification tasks to enhance the performance of acoustic level modeling [174] (please refer to Section 2.2.5 for a full description of feature normalization techniques). In all these studies, age and gender were recognized jointly within a broad set of age classes corresponding to children, young adults, adults, and seniors. The gender and age-groups were not considered for the children's class, and we do not know the significance of different frequency bands for age group or gender identification using children's speech.

7.2 Experimental Objectives

This chapter presents the results of experiments on gender identification (GI) and Age-group identification (Age-ID) from children's speech, and is organized as follows: Section 7.3 describes the speech data used in all experiments. Our GI and Age-ID systems are described in Section 7.4, and our experiments and results are presented in Section 7.5. First, we describe a study of the information's utility in different frequency bands for children's GI and Age-ID in Sections 7.5.1 and 7.5.2, respectively. Next, in Section 7.5.3, we explore the effect of using age-independent and age-dependent gender modeling and the use of GMM-UBM, GMM-SVM and i-vector approaches. We also analyse the effect on GI of voice breaking for children in the oldest group. Further, we present the effect of employing intersession variability modeling in Section 7.5.3.3. Similar sets of experiments are also conducted using full and restricted bandwidth speech for Age-ID task, and the results are presented in Section 7.5.4. Finally, in Section 7.5.5 and Section 7.5.6, the GI and age-ID performances, respectively, by human listeners and machines are compared.

7.3 Data Description

7.3.1 Gender identification

The OGI Kids' Speech corpus [141] is a collection of spontaneous speech and those from readings recorded at the Northwest Regional School District near Portland, Oregon. A full description of this corpus is available in Section 3.2.1. 3 different gender-balanced test sets from the OGI data are used in the experiments presented in this chapter.

TS1: To investigate the effect of different frequency bands on GI performance for general children's speech, 687 speakers were chosen randomly (from kindergarten to 10th grade).

TS2: To investigate the effect of different frequency bands and using age dependent models on GI performance for speech from children of different ages, 3 different age groups were selected, each containing 76 speakers. These are as follows:

AG1: Kindergarten to 3rd grade (5-9 years old),

AG2: 4^{th} to 7^{th} grade (9-13 years old), and

AG3: 8^{th} to 10^{th} grade (13-16 years old).

TS3: To investigate the effect on gender identification of the voice breaking for male speakers going through puberty, the data in the AG3 group was split into three sub-sets, denoted as "boys broken," "boys unbroken," and "girls." Each sub-set contained data from 18 speakers for training. For testing, all 191 speakers from AG3 (both male and female) were used.

7.3.2 Age-group identification

The OGI kids corpus is also used for investigation of the age-group identification task. In this study, 766 speakers were chosen randomly for testing, and the remaining 334 for training. The age groups are the same as specified for TS2 in the previous Section. The individual age groups, AG1, AG2, and AG3 contained 290, 285, and 191 test speakers, respectively.

7.4 Age-group and Gender Identification Systems

7.4.1 Signal analysis

Feature extraction was performed as follows: Periods of silence were discarded using an energy-based SAD. The speech was then segmented into 20-ms frames (10-ms overlap), and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying an FFT, is passed to a bank of 24 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 8000 Hz. Table 6.3.1 shows the center frequency of each filter (the cut-off frequencies of a filter are the center frequencies of the adjacent filters). To investigate the effect of different frequency regions on GI and Age-ID performance, experiments were conducted using frequency-band limited speech data comprising the outputs of groups of 4 adjacent filters. We considered 21 overlapping sub-bands, in which the N^{th} sub-band comprises the outputs of filters N to $N+3$ (N=1 to 21). Each set of 4 filter outputs was transformed to 4 MFCCs plus 4 delta and 4 delta-delta parameters, and feature warping [2] was applied (a full description of feature warping method is included in Section 2.2.5.3). For the full bandwidth experiments the outputs of all 24 filters were transformed into 19 MFCCs plus 19 deltas and 19 delta-deltas.

7.4.2 Modeling

Our Age-ID and GI systems are based on the GMM-UBM [80, 83, 169], GMM-SVM [83] and i-vector [87, 92] methods; please refer to Section 2.3.3, Section 2.4.1, and Section 2.4.2.2, respectively for a full description of these systems.

In the GMM-UBM approach, a UBM is built using all utterances from 418 and 334 speakers, for GI and Age-ID, respectively. For the GI the age-independent gender models are obtained by MAP adaptation (adapting means only) of the UBM, using the gender-specific enrollment data (there is no overlap between the training and test sets). The result is one UBM and 2 gender-dependent GMMs. To investigate the effect of using age-dependent gender models on GI performance, age-dependent models are obtained by MAP adaption of the UBM using the age and gender-specific enrollment data (there is no overlap between the training and test sets). For Age-ID, the gender-independent age group models are obtained by MAP adaptation (adapting the means only) of the UBM, using the age-group-specific training data. The result is 1 UBM and 3 age-group GMMs. To investigate the effect of using gender-dependent age-group models on Age-ID performance, gender-dependent models are obtained by MAP adaptation of the UBM, using the gender and age-group-specific training data.

In our GMM-SVM system, the speech data from each gender and age group category were used to estimate the parameters of a GMM by MAP adaptation of the UBM. The adapted GMM mean vectors are then concatenated into a super-vector (as described in Section 2.4.1.2) [83], and the gender/age classes are assumed to be linearly separable in the super-vector space. The super-vectors are used to build one SVM for each gender class by treating that gender/age class as the 'target' class and the others as the 'background' class.

The same data that was used to train the UBM was used to train the "T-matrix" in the i-vector systems. In the total variability modeling approach, i-vectors are the low-dimensional representation of an audio recording that can be used for classification and estimation purposes. A full description of the i-vector frame work is included in Section 2.4.2.

7.5 Experimental Results and Discussion

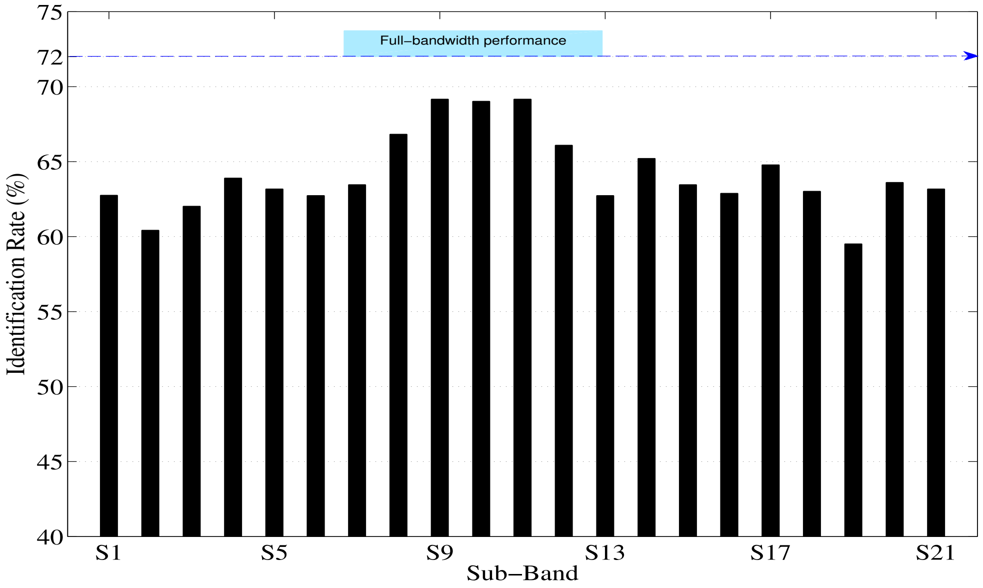
7.5.1 Sub-band based gender identification for children's speech

In this section, we study the effect of different sub-bands on GI performance for children's speech. Experiments are conducted separately on 21 sub-bands, each consisting of four consecutive channels (see Section 7.4.1 for more details), and using the age-dependent GMM-UBM system. For each sub-band, three age-dependent models of each gender are trained. The models have 64 mixture components, which were found to be adequate for these 12 dimensional sub-band features.

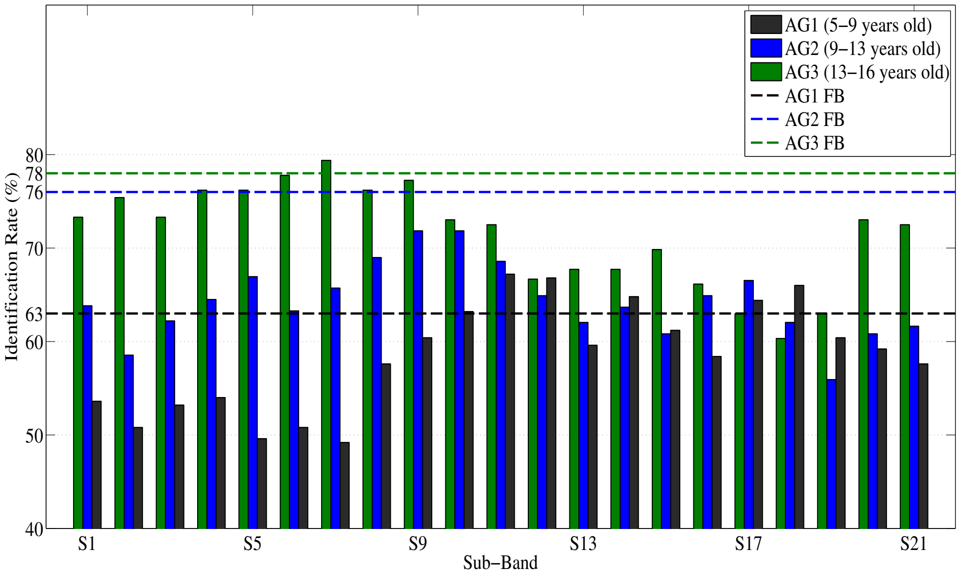
Figure 7.5.1(a) presents the average results across all age groups. The most useful sub-bands for GI are from 8 to 12 (frequency range 0.9 kHz–2.6 kHz). This corresponds to the location of the second formant for vowels, which was also found to provide the best GI performance for adult speakers in [175].

Figure 7.5.1(b) shows the performance for each age group. For AG3 (i.e., the oldest children), the frequency sub-bands up to 9 (frequencies up to 1.8 kHz) and above 19 (frequencies above 3.8 kHz) provide somewhat similar performances of around 75%, while the middle frequency sub-bands give lower performance. For AG2, the performance does not vary largely across the frequency bands. The peak performance is achieved at sub-bands 9 and 10 (frequency range 1.0 kHz–2.1 kHz). For AG1, the performance is close to chance for sub-bands up to 7 (frequencies up to 1.4 kHz) and then increases to around 65% for sub-band 11 and 12 and stays fluctuating around 60% for the remaining higher sub-bands.

It may be that the insignificance of sub-bands up to 11 for young children, and their increasing utility as the age of the child increases is due to greater and more consistent vocal effort in older children. Our hypothesis is that for the low frequency bands (up to sub-band 11) the poor performance of the young children (speakers in AG1) is due to the wide spacing of their pitch harmonics, and that as the children become older and their pitch lowers, the harmonics come closer together, and the problem diminishes. The problem with the widely-spaced pitch harmonic is that they may not captured by the initial triangle filters, as these are narrow filters. The Mel scale is approximately linear up to 1000Hz; a critical band is about 100Hz, so the first 10 triangular filters are approximately of the same bandwidth (with an Equivalent Rectangular Bandwidth (ERB) of about 100Hz). Thus, if the effect is due to pitch harmonics, one would expect it to start diminishing when the bands are no longer influenced by these narrow filters. The first band that is not influenced by the narrow triangular filters is band 11. This is the band in which the performances of the different ages start coming together. Therefore, this finding supports the hypothesis that the effect is due to the narrow bands and the high fundamental frequency of the youngest children. Based on this assumption, we would expect the problem to cease as the triangular filters become broader. The solution, according to Shweta Ghai's work



(a)



(b)

Figure 7.5.1: The effect of different frequency sub-bands on gender identification: average over all age groups (a) and for each of the three children age groups (b).

on ASR [7], is to broaden the width of the low frequency filters so that the ERB or the width of the triangles' base is 300Hz. Figure 7.5.2 shows the original and modified filter bank used in Shweta's work.

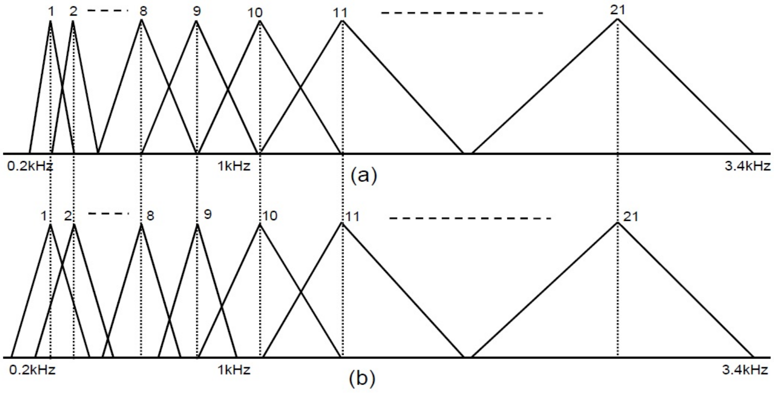


Figure 7.5.2: Structures of the Mel filterbank (a) Default (b) Modified. In the modified filterbank the bandwidth of all filters having center frequency below some particular frequency value (say 1 kHz) are modified to have a constant value whereas those of the other filters remain unchanged (taken from [7]).

This solution is proposed but due to lack of time it is not applied for gender identification task.

7.5.2 Sub-band based Age-ID for children's speech

In this section, we study the effect of different sub-bands on Age-ID performance for children's speech. Experiments are conducted separately on 21 sub-bands, each consisting of four consecutive channels (see Section 7.4.1 for more details), and using the gender independent GMM-UBM system. For each sub-band, 3 gender-independent (based on the results of full-band experiments) age-group models are trained, corresponding to AG1, AG2 and AG3. Corresponding to AG1, AG2, and AG3, the models have 64 mixture components, which were found to be adequate for these 12 dimensional sub-band features.

Figure 7.5.3 presents the average Age-ID results as a function of frequency sub-band. It is evident that the performance, even when using a narrow frequency region, is in most cases well above chance. The best performance achieved by using sub-bands 13 to 16 and sub-bands 18 to 21 represents the least useful bands for age identification. Figure 7.5.4 contrasts the usefulness of sub-bands for Age-ID and Gender-ID. The figure was obtained by normalising the data in Figure 7.5.3 so that the sum of the values over all of the sub-bands is 1. The same procedure was then applied to the corresponding sub-band results on Gender-ID presented in [39] and Section 7.5.1 (please refer to Figure 7.5.3 (a)). The normalised Age-ID results were then subtracted from the normalised Gender-ID results to obtain Figure 7.5.4 (similar procedure is described in [36]). Thus, negative regions in Figure 7.5.4 indicate sub-bands which are more useful for Age-ID while positive values indicate sub-bands that are useful for Gender-ID. The results indicate that the most useful sub-bands for Age-ID, in comparison to Gender-ID, are the sub-bands 3 and 4 (281 Hz to 625 Hz), and from 13 to 16 (1.62 kHz to 3 kHz). Thus, while Gender-ID appears to make use of similar information to speaker recognition, Age-ID is more similar in these respects to speech recognition or accent ID [36] (please refer to Section 5.5).

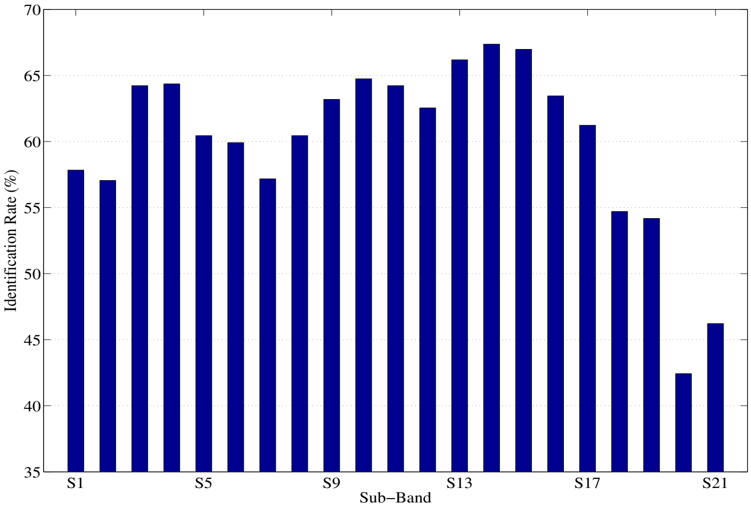


Figure 7.5.3: The effect of different frequency sub-bands on Age-ID, average identification rate over all age groups. Using full-bandwidth speech signals the performance was 82%.

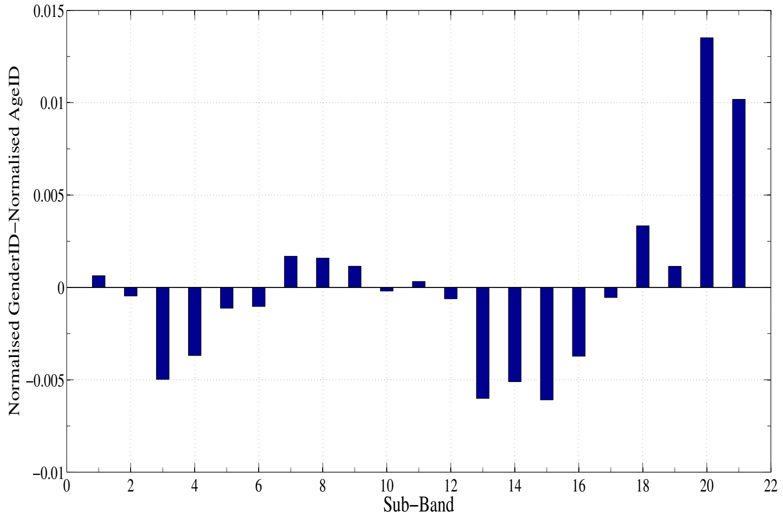


Figure 7.5.4: The difference between the normalized Gender-ID and Age-ID performance for frequency sub-bands.

7.5.3 Full-bandwidth gender identification for children's speech

This section demonstrates the effect of using different modelling approaches. Experiments were performed using full-bandwidth speech.

7.5.3.1 Age-independent modelling

First, we demonstrate the effects of employing the generative GMM-UBM, discriminative GMM-SVM, and i-vector systems when using age-independent modelling. For each of the systems, we performed experiments using different numbers of mixture components. The best results were obtained when using 1024, 512, and 256 mixture components for the GMM-UBM, GMM-SVM, and i-vector systems, respectively, and these are presented in Table 7.5.1. It can be seen that the GMM-SVM system considerably outperforms the GMM-UBM and the i-vector systems.

7.5.3.2 Age-dependent modelling

This section demonstrates the effect of using age-dependent modelling, in which all training data is split into three age groups (as described in Section 7.3), and a model is created for each

Table 7.5.1: Gender identification performance obtained by the age-independent GMM-UBM, GMM-SVM, and i-vector systems.

System	GI rate (%)
GMM-UBM (age independent)	67.39
GMM-SVM (age independent)	**77.44**
i-vector (age independent)	74.26

age group. During recognition, models corresponding to the age of the speaker of the testing utterance were used. The experiments' results are presented in Table 7.5.2. It is evident that the performance of the GMM-UBM, GMM-SVM and i-vector systems improved by 4.23%, 1.74%, and 0.28%, respectively, compared to corresponding age-independent systems. The small improvement in the i-vector system could be due to the importance of training data amount on the performance of factor-analysis-based recognition systems. It is true that by separating the data with respect to the age of the speaker, the amount of complexity is reduced; however, simultaneously, the training data amount for each particular class is reduced dramatically by training the age-dependent gender models, compared to the age-independent models.

Table 7.5.2: Gender identification performance obtained by the age-dependent GMM-UBM, GMM-SVM, and i-vector systems.

System	GI rate (%)
GMM-UBM (age dependent)	71.76
GMM-SVM (age dependent)	**79.18**
i-vector (age dependent)	72.54

We then analysed the results obtained by the age-dependent GMM-UBM and GMM-SVM systems for each age group. These are presented in Table 7.5.3, in which "B," "G," and "Av" denotes boys, girls, and average. This Table shows that the boys' performance is consistently poorer (except for AG3 using GMM-SVM system) than girls'. One would expect the GI performance to be lowest for youngest children, i.e. AG1, and to improve as the age increases. Indeed, one can see that the identification rate achieved by each system for AG2 is considerably higher than for AG1 – the performance increase is 12.81% for the GMM-UBM and 7.34% for the GMM-SVM system. However, the performance is unexpectedly low for AG3 (i.e., the oldest children); compared to AG2, the performance improves only by 2.52% for the GMM-UBM and decreases by 8.23% for the GMM-SVM system. This may be related to the fact that the boys in AG3 fall into two subsets, according to whether or not their voices have broken as a consequence of puberty. Results suggest that the GMM-UBM system is better able to accommodate this issue than the GMM-SVM system.

As mentioned earlier in Section 2.7.4, the timing and tempo of puberty vary widely, even among healthy children, but there are several studies which tried to estimate the distribution for age of pubertal growth for girls and boys, separately [139] (for more details please refer to Section 2.7.4).

Table 7.5.3: Gender identification performance (in %) obtained by the age-dependent GMM-UBM and GMM-SVM systems for each age group.

Age	GMM-UBM			GMM-SVM		
group	B	G	Av	B	G	Av
AG1	40.00	90.09	63.20	73.33	80.86	76.80
AG2	69.67	82.25	76.01	79.50	88.70	84.14
AG3	70.00	88.52	78.53	77.69	72.13	75.91

We further analysed the effect of voice breaking in AG3. This was performed using the GMM-UBM system. Table 7.5.4 (a) shows the GI confusion matrix for AG3 when using a single model for boys and a single model for girls. One can observe that there is a high confusion regarding the gender of boys being recognised as girls. We speculate that this is because the model for "boys" covers broken and unbroken voices, and consequently, some boys whose voices have not broken may achieve a better match with the "girls" speech model. These results, and the fact that changes in the voice coinciding with puberty is prominent mainly in boys rather than in girls, motivated us to split the data of boys in AG3 into two separate classes: boys whose voice had broken, denoted as BB, and whose voice remained unbroken, denoted as BU. Categorizing the boys' data into these two classes was performed by a human listener. Some further details regarding the resulting training and testing data are described in Section 7.3.1, where it is denoted as TS3. Because this resulted in reduced amounts of training data for each class, the GMM-UBM system for each of the AG3 gender sub-groups consisted of 128 mixture components. When using the 3 gender sub-group models, the average GI rate for AG3 was 87.43%. This is an improvement of 8.9% from 78.53%, achieved by the system using two gender models (and each consisting of 256 mixture components) as presented in Table 7.5.3. Table 7.5.4

(b) presents the confusion matrix corresponding to this experiment. The amount of gender confusion from boys to girls is reduced to zero. The amount of confusion from girls to boys with unbroken voices is much larger than to boys with broken voices, which is expected. The performance for girls decreased when boys were divided into two categories, and this could be due to the fact that some girls' voices in AG3 are broken, as a matter of puberty, and they are confused with boys' with unbroken voices. These results suggest that by dividing girls into girls with broken and girls with unbroken voices we may gain performance improvement.

Table 7.5.4: Confusion matrix for gender identification (in %) for age group AG3 when using for boys a single model (a) and two separate, broken B_B and unbroken B_U, models (b).

(a)

	B	G
B	71.5	28.5
G	11.5	88.5

(b)

	B_B	B_U	G
B_B	96.2	3.8	0
B_U	5.8	94.2	0
G	6.5	23.0	70.5

7.5.3.3 Effect of intersession variability modelling

We also investigated the effect of intersession variability (ISV) compensation on model domain (readers are referred to Section 2.4.3 for a full description). These experiments were performed using the GMM-UBM system, and both age-independent and age-dependent modelling. Both systems achieved only small performance improvements when ISV modelling was applied, specifically, the age-independent system improved from 67.39% to 69.29%, and the age-dependent system improved from 71.76% to 72.81%.

7.5.4 Age-ID using full/restricted bandwidth speech

This section presents the Age-ID results that are obtained using the GMM-UBM, GMM-SVM, and i-vector based systems described in section 7.4.2. Experiments were performed using full-bandwidth (FB) and band-limited speech (BL). The band-limited case includes frequencies up to 5.5 kHz, which corresponds to the frequency region covered by all sub-bands except sub-bands 18 to 21 (Figure 7.5.3 suggests this exclusion of sub-bands).

We first study the effect of using gender-dependent and independent age-group modeling, using the GMM-UBM system. The results of this study are shown in the first two rows of Table 7.5.5. For Age-ID, gender-independent modelling gives better results than gender-dependent modeling. This could be due to two phenomena; the first is the smaller amount of training data (half) available for gender-dependent age modelling compared to gender-independent age modelling, and the second is that dividing AG1 speakers with respect to their gender does not provide any benefit because for AG1 there are limited acoustic cues for GID (Table 7.5.3). It is probably a trade-off in that any benefit from gender-dependent modelling in AG1 is off-set by the effect of the small training sets. Based on these results, subsequent experiments use gender-independent modelling.

Then, we demonstrate the effects of employing the discriminative GMM-SVM and i-vector systems when using gender-independent modelling. For each of the systems, we performed experiments using different numbers of mixture components. The best results were obtained when using 1024, 512 and 256 mixture components for the GMM-UBM, GMM-SVM, and i-vector system, respectively, and these are presented in Table 7.5.5. For the i-vector system, we performed experiments using different numbers of dimensions for training the total variability matrix. The best results were obtained using 400 dimensions for the T matrix. The i-vector system evidently outperforms the GMM-UBM and GMM-SVM systems, especially when band-limited speech is used.

Table 7.5.5: Age-ID recognition rate (in %) obtained by the gender-independent GMM-UBM, GMM-SVM and i-vector systems and gender-dependent GMM-UBM system.

System	Age-ID rate (%)	
	Full-bandwidth	Band-limited
GMM-UBM (gender dep.)	71.76	-
GMM-UBM (gender indep.)	82.01	84.07
GMM-SVM (gender indep.)	79.77	-
i-vector (gender indep.)	**82.62**	**85.77**

Table 7.5.6 depicts a confusion matrix obtained by the i-vector system using band-limited speech. Each row corresponds to a grade and shows the percentages of children in that grade who were classified as being in AG1, AG2, and AG3. The dotted lines indicate the boundaries of AG1, AG2, and AG3. The top and bottom halves of the table correspond to male and female

Table 7.5.6: Confusion matrix for age identification (in %) for three age groups, obtained by the i-vector system using band-limited speech.

	Model-index		
Grade-index	AG1 (%)	AG2 (%)	AG3 (%)
Male			
k	100	0	0
1^{st}	100	0	0
2^{nd}	97.43	2.56	0
3^{rd}	85.71	10.20	4.08
4^{th}	33.33	60.60	6.06
5^{th}	8.57	82.85	8.57
6^{th}	6.97	81.39	11.62
7^{th}	0	54.83	45.16
8^{th}	0	0	100
9^{th}	2.17	6.52	91.30
10^{th}	0	0	100
Female			
k	100	0	0
1^{st}	100	0	0
2^{nd}	97.87	2.12	0
3^{rd}	92.10	7.89	0
4^{th}	38.70	61.29	0
5^{th}	11.42	82.85	5.71
6^{th}	10.00	80.00	10.00
7^{th}	2.70	72.97	24.32
8^{th}	0	29.03	70.96
9^{th}	5.00	20.00	75.00
10^{th}	0	30.00	70.00

speakers, respectively. The table shows similar characteristics for boys and girls up to the 7^{th} grade, with the majority of errors near age-group boundaries. At the boundary between AG1 and AG2, 10% of 3^{rd} grade boys (AG1) are incorrectly classified as AG2, and 33% of 4^{th} grade boys (AG2) are incorrectly classified as AG1. For girls, the corresponding figures are 8% and 39%. For 7^{th} grade, (AG2) 45% of boys and 24% of girls are classified as being in AG3, while for 8^{th} grade, (AG3) 29% of female speakers are classified as AG2, but none of the boys are misclassified. The inconsistency between the results for boys and girls at the AG2-AG3 boundary may be because AG3 contains speech from a number of boys whose voices have broken. It may be that gender-dependent modelling is needed for AG3, even though it is not advantageous overall, or that, as in the case of gender identification [39], it is necessary to build separate models for AG3 boys whose voices have or have not broken.

7.5.5 Human GI for children's speech

In addition to the computer GI experiments presented in the previous sections, we also performed experiments on GI by human listeners. The test set consisted of the same 687 test utterances used in the computer GI experiments. Twenty listeners participated in the experimental evaluations. Each participant listened to 34 utterances on average and there is no overlap between utterances listened to. The length of each utterance was 10 seconds. All human listening tests were performed in a quiet room using the same PC and headphones.

The GI rates for each age group achieved by human listeners are presented in Table 7.5.7. The average performance of all age groups was 66.96%.

Table 7.5.7: Gender identification performance for each age group obtained by human listeners.

Age group	Human GI rate (%)
AG1	60.48
AG2	70.49
AG3	70.90

As is depicted in Table 7.5.7, as with computers, humans are not as good as was initially expected for the GID task, particularly when using children from AG3. To find out how much variation in performance is there between the listeners the ANOVA test was applied. We wanted to investigate whether the differences in performance between the AG2 and AG3 are statistically significance in the view of the differences between subjects. For performing ANOVA test the built in Matlab function (anova1) was used. This function compares the means of two observations, observations from performances obtained by listeners for test utterances of AG2 and AG3. The function returns the p-value under the null hypothesis that all samples are drown from populations with the same mean [176]. The obtained p-value for this test was 0.1974, which is relatively large and shows that the there is no reason to conclude that means differ. So we can say that the accuracy improvement for AG3, compared to AG2, is relatively small - only 0.41% and negligible.

7.5.6 Human Age-ID for children's speech

In addition to the computer Age-ID experiments presented in the previous sections, we also performed experiments on Age-ID by human listeners. The test set consisted of the same 766 test utterances used in the computer Age-ID experiments. Twenty listeners participated in the experimental evaluations. Each participant listened to 38 utterances on average and there is no overlap between utterances listened to. The length of each utterance was 10 seconds. All human listening tests were performed in a quiet room using the same PC and high-quality headphones.

The Age-ID rates for each age group achieved by human listeners are presented in Table 7.5.8. This table suggest that the main confusion is came from the test utterances of AG2. Only 50.8 % of tests from AG2 are correctly identified and the rest are confused with AG1 and AG2. The confusion between AG1 and AG3 is small and only 1.8 % and 3.8 % of the test utterances form AG1 and AG3 are miss identified as AG3 and AG1, respectively. The average performance of all age groups was 67.54%.

Table 7.5.8: Confusion matrix for age identification (in %) for three age groups, obtained by human listeners.

	Model-index		
Test-index	AG1	AG2	AG3
AG1	**81.2**	16.9	1.8
AG2	25.5	**50.8**	23.6
AG3	3.8	24.4	**71.7**

As depicted in Table 7.5.8 the worst performance for human listeners obtained for children in AG2. It was expected as this age-group has a border with both AG1 and AG2, which will results in poor performance for this age-group.

7.6 Summary

This chapter presents the results of experiments in Age-ID and GI for children's speech using the OGI kids' speech corpus.

A study of the different narrow frequency bands' utility has shown that the frequency region 0.9–2.6 kHz is the most useful by average of all age groups for GI. The separate analysis of the results for each of the three age groups show that the performance trend is different for each age group. For the AG3 children (13-16 years), the frequencies up to 1.8 kHz and above 3.8 kHz provide the best performance, which is around 75%. For AG2 children (9-13 years) performance does not vary largely across the frequency bands. For AG1 children (5-9 years), the performance is close to chance, up to 1.4 kHz, and then increases to 65% for sub-bands around 1.8 kHz, fluctuating around 60% for frequencies above 2.3 kHz. Figure 7.5.1 (b) shows that the main differences between younger and older children are for bands up to sub-band 11. Below sub-band 11, the performance decreases with age, and this may be due to the gap between harmonics increasing as age decreases, and because the triangle filters are narrow at low frequency. Above sub-band 11, the difference is less clear across different ages. For the two highest bands, performance is much better for the older children. Above sub-band 11, and below sub-band 20, the sub-bands produce similar performance across all ages.

The effect of using age-dependent gender modelling and the GMM-UBM, GMM-SVM and i-vector techniques was examined using the full-bandwidth experiments. The age-independent GMM-SVM system outperformed the GMM-UBM and i-vector systems by nearly 10% and 3%, respectively. The age-dependent gender models gave 4.23%, 1.74%, and 0.28% GI improvement in the case of the GMM-UBM, GMM-SVM and i-vector systems, respectively. The full-bandwidth results for each age group were analysed, which showed unexpectedly low performance for the AG3 children. An investigation confirmed that this was due to the fact that the "boys" category in AG3 includes both boys with broken and unbroken voices, depending on whether or not the child had entered puberty. Consequently, speech from boys whose voices have not broken may achieve a better match with the "girls" acoustic model. The data of boys in AG3 was divided into 2 separate groups: Boys with broken and unbroken voices. The use of the 3 gender classes provided a GI rate of 87.34% for the AG3 children, which was an improvement of 8.81% from using only two gender classes. The application of intersession variability compensation was explored, but experiments showed only little improvement. Human GI experiments were also conducted, and the average performance for all age groups was 66.96%, which is lower than the performance achieved by machine.

Our results for Age-ID based on narrow frequency sub-bands indicate that the performance, even for narrow frequency regions, is in most cases well above chance. The best performance seems to be obtained when using sub-bands 13 to 15. Moreover, a comparison of useful bands for Age-ID and Gender-ID shows that most of the useful information for Age-ID is in similar regions of the spectrum to those that are useful for accent ID. This result suggests that removing higher parts of the spectrum will improve Age-ID performance. Hence, we compared Age-ID performance for full bandwidth (up to 8 kHz) and restricted bandwidth (up to 5530 Hz). As expected, performance is improved for both the GMM-UBM and i-vector systems when band-limited speech is used. The best Age-ID performance is 87.55%, obtained from the i-vector system applied to band-limited speech. Further analysis of the results from the best system shows that Age-ID for young children, both male and female speakers, is a relatively easy task. The main confusion arises with male and female speakers who belong to the 4^{th} and 7^{th} grades. These grades are at the boundaries of AG2. For example, 33.33% and 38.70% of 4^{th} grade boys and girls are misidentified as belonging to AG1, respectively. It is also evident that for AG3, Age-ID for girls is more challenging than for boys from same age group.

Human experiments for both GI and Age-ID show that computers outperform humans (untrained listeners) considerably. For GI, when using children from AG3, the accuracy of humans is unexpectedly low, as with computers, which is due to the inclusion of boys and girls with broken voices in this particular age group.

"Yesterday I was clever, so I wanted to change the world. Today I am wise, so I am changing myself."

Jalāl ad-Dīn Muhammad Rūmī

8 Conclusion

The major contributions of this thesis could be summarized as follows: First it proposes a new analysis for investigation of the effectiveness of different parts of the speech spectrum for speaker and accent identification, using adult speech. The next important contribution is the study of the utility of state of the art speaker recognition techniques for children's speech, and comparison of the regions of the spectrum that are most important for speaker recognition for children and adults. The third major contribution is a investigation of the application of speaker identification to the problem of identifying child in a simulated class and school of children. The forth major contribution is a study of gender identification systems for children's speech, by humans and machines, and study the utilities of different parts of the speech spectrum. Lastly, age-group identification, using children's speech, is investigated by humans and computers, and useful bands for this task are identified.

8.1 Summary of Results

In this research we first focused on the speaker recognition task, using adult's speech and NIST SRE plan, to obtain a baseline system which has a similar performance compare to the state of the art speaker recognition systems. Then, the results of an experimental study investigating the effect of frequency sub-bands on regional AID and SID performance on the ABI-1 corpus are presented. The AID and SID systems are based on Gaussian mixture modelling. The SID experiments show up to 100% accuracy when using the full 11.025 kHz bandwidth. The experiments using isolated narrow sub-bands show that the regions (0–0.77 kHz) and (3.40–11.02 kHz) are the most useful for SID, while those in the region (0.34–3.44 kHz) are best for AID. AID experiments are also performed with intersession variability compensation, which provides the biggest performance gain in the (2.23–5.25 kHz) region. The best AID performance of 76.76% is obtained when using an i-vector system and band-pass filtered (0.23–5.5 kHz) speech.

Although speaker verification is an established area of speech technology, previous studies have been restricted to adult speech. By having an appropriate level of understanding from adult speech and its relevant applications, we start assessing various classification tasks using information in children's speech. In a very first set of experiments we presents results on SR for children's speech, using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vector SR systems. Regions of the spectrum containing important speaker information for children are identified by conducting SR experiments over 21 frequency bands. As for adults, the spectrum can be split into four regions, with the first (containing primary vocal tract resonance information) and third (corresponding to high frequency speech sounds) being most useful for SR. However, the frequencies at which these regions occur are from 11% to 38% higher for children. It is also noted that sub-band SR rates are lower for younger children. In addition results are presented of SR experiments to identify a child in a class (30 children, similar age) and school (288 children, varying ages). Class performance depends on age, with accuracy varying from 90% for young children to 99% for older children. The identification rate achieved for a child in a school is 81% and 87.15%, when using GMM-UBM and i-vector systems, respectively. In addition a contemporary GMM-based speaker verification system, using MFCC features and maximum score normalization, is applied to adult and child (for experiments using child speech PF-STAR corpus is used) speech at various bandwidths using comparable test and training material, to enable us have a comparison between SR from adult and child speech. The results show that the EER for child speech is almost four times greater than that for adults. A study of the effect of bandwidth on EER shows that for adult speaker verification, the spectrum can be conveniently partitioned into three frequency bands: up to 3.5-4kHz, which contains individual differences in the part of the spectrum due to primary vocal tract resonances, the region between 4kHz and 6kHz, which contains further speaker-specific information and gives a significant reduction in EER, and the region above 6kHz. These finding are consistent with previous research [31]. For young children's speech a similar pattern emerges, but with each region shifted to higher frequency values.

Similar experiments are also conducted for GI using children's speech. The results are obtained by using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vectors systems. As for SR, regions of the spectrum containing important gender information for children are identified by conducting GI experiments over 21 frequency sub-bands. Results show that the frequencies below 1.8 kHz and above 3.8 kHz are most useful for GI for older children, while the frequencies above 1.4 kHz are most useful for the youngest children. The effect of using age-independent and age-dependent gender modelling (including the effects of puberty on boys voices) is explored. The application of intersession variability compensation is explored but experiments showed only little improvement. Experiments on human GI were also conducted and the results show that the humans do not achieve the performance of the machine.

Lastly, the results on Age-ID for children's speech, using the OGI Kids corpus and GMM-UBM, GMM-SVM and i-vector systems, are presented. Regions of the spectrum containing important age information for children are identified by conducting Age-ID experiments over 21 frequency sub-bands. Results show that the frequencies above 5.5 kHz are least useful for Age-ID. The effect of using gender-independent and gender-dependent age-group modelling is explored. The GMM-UBM and i-vector systems considerably outperform the GMM-SVM system. The best Age-ID performance of 85.77% is obtained by the i-vector system applied to band-limited speech to 5.5 kHz. Experiments on human Age-ID were also conducted and the results show that the machine outperforms the humans, by 18.23%.

8.2 Future Research Directions

The majority of this work focused on a study of state of the art detection technologies to recognize children's identity, gender and age-group, using short speech samples. State-of-the-art detection techniques are changing all the time, and new approaches are constantly emerging from language and speaker recognition fields. The application of the most recent techniques to the problem of speaker characterization, using children's speech, is an interesting ongoing area of research.

Because of extensive research effort on automatic speech recognition, a variety of normalization techniques are proposed by researchers. One interesting research question could be: are the normalization techniques from speech recognition useful or not for detection tasks? The answer to this question strongly depends on the application, for example pitch normalization could be useful for gender identification but it may not be good for age identification, as pitch information is likely to be useful for age identification and therefore normalizing pitch is a bad idea in case of age identification. This hypothesis is based on our findings from Figure 7.5.1 (b) (where the low frequencies seems to be of little use for younger children, when using the typical design of filter bank filters) and Figure 7.5.3.

Our current findings on gender identification from children's speech showed unexpectedly low performance for the AG3 children. An investigation confirmed that this was due to the fact that the 'boys' category in AG3 includes both boys with broken and unbroken voices, depending on whether or not the child has entered puberty. Consequently, speech from boys whose voices have not broken may achieve a better match with the 'girls' acoustic model. The data of boys in AG3 were divided into two separate groups, boys with broken and unbroken voices. Table 7.5.4 shows the gender confusion from boys to girls is reduced to zero, but confusion from girls to boys with unbroken voices is much larger than to boys with broken voices. The performance for the girls has gone down when the boys are divided into two categories. This could be due to the fact that voices of some of the girls in AG3 are affected by puberty, and they are confused with boys with unbroken voices. These results suggest that by dividing girls into different categories according to effects of puberty may result in performance improvements.

In addition to the computer age and gender identification techniques presented in the thesis, we also performed experiments on age and gender identification by human listeners. For human experiments, the human listeners are chosen from research students, mainly single and only few participant who have children (and who could therefore be considered as expert/trained listeners). Familiarity with children could affect the human listener's performance. Rerunning the human experiments with the trained listeners seems to be an interesting experiment. School teachers, nurses, and even mothers could be considered as trained listeners.

For age-group identification from children speech, the performance could potentially be improved by employment of phonotactic approaches, which are based on the occurrence of different phone sequences. This type of approach seems practically useful for age-group identification using children's speech, as it is intended to capture differences which are caused by rapid language development of children. Finding a relationship between age and language skill is also an interesting area of research, as they clearly are correlated. For example 36 months child have access only to approximately 900-1000 words, but as they grow up this number increases dramatically. In addition young children often repeat words, phrases, and syllables. By obtaining scores using phonotactic system, next step is the fusion of these scores with the scores obtained from from acoustic systems.

All these detection technologies then could be combined to form a automated tutor system for schools and classrooms, so children could login in to the system using her/his speech samples and based on a user's gender and age-group, relevant teaching materials will be provided to them. To make an automatic tutor system more powerful and interactive, having an automatic speech recognition system, based on children speech, is essential. Hence studying the utility of state of the art speech recognition techniques (currently deep neural networks take the lead) for children's speech is an important area of research for the future. The findings of the thesis could be effectively used during the design of automatic speech recognition system for children, for example speaker, gender, and age-group identifications from children's could be used for model selection in speech recognition.

"Silence is the language of god, all else is poor translation."
Jalāl ad-Dīn Muhammad Rūmī

A
Appendix

A.1 List of Publications

A.1.1 Articles in international journals

1. **S. Safavi**, A.Hanani, M. Russell, P. Jancovic and M. Carey, "Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification.", IEEE Signal Process. Lett. 19: 829-832,2012.

A.1.2 Articles in international conferences

1. **S. Safavi**, M. Najafian, A.Hanani, M. Russell, P. Jancovic and M. Carey , "Speaker recognition for children's speech", Proc. Interspeech 2012, Portland, USA, 2012.

2. **S. Safavi**, P. Jancovic, M. Russell and M. Carey, "Identification of gender from children's speech by computers and humans.", Proc. Interspeech 2013, Lyon, France, 2440-2444, 2013.

3. **S. Safavi**, M. Russell, and P. Jancovic, "Identification of age-group from children's speech by computers and humans.", Proc. Interspeech 2014, Singapore, 2014.

4. **S. Safavi**, M. Najafian, A. Hanani, M. Russell and P. Jancovic,"Comparison of Speaker Verification Performance for Adult and Child Speech.", WOCCI 2014.

"What you seek is seeking you."

Jalāl ad-Dīn Muhammad Rūmī

B Appendix

B.1 Distance Measure

Suppose there are two utterances, *utt* and *utt*′. GMMs, G_{utt} and $G_{utt'}$, are adapted from a UBM using map adaptation (means only). The distance between these two utterances could be calculated using Euclidean distance between the scaled GMM super-vectors s and s',

$$d(s, s') = \frac{1}{2} \sum_{i=1}^{C} w_i (s_i - s'_i) \Sigma_i^{-1} (s_i - s'_i) \qquad \text{(B.1)}$$

where s and s' are the super-vectors obtained from the adapted means of utterance *utt* and *utt*′, respectively. w_i and Σ_i are the i^{th} UBM mixture weights and diagonal covariance matrix, respectively, and s_i corresponds to the mean of Gaussian i of the speaker GMM. The derived linear kernel is defined as [83],

$$K_{linear}(s, s') = \sum_{i=1}^{C} w_i s_i \Sigma^{-1} s'_i = \sum_{i=1}^{C} (\sqrt{w_i} \Sigma^{-\frac{1}{2}} s_i)(\sqrt{w_i} \Sigma^{-\frac{1}{2}} s'_i)^t = b(x)^t b(y) \qquad \text{(B.2)}$$

"The most exciting phrase to hear in science, the one that heralds the most discoveries, is not 'Eureka!' (I found it!) but 'That's funny...' "

Isaac Asimov

C Appendix

C.1 Joint Factor Analysis

In 2005, Kenny [84] gave a report on the theory and algorithm needed to carry out a Joint Factor Analysis (JFA) model of speaker and session variability in a training set of multiple recording sessions per speaker. Most of the variance in the super-vector population is assumed to be accounted for by a small number of hidden variables, which they referred to as speaker (or any other type of speaker's characteristics) and channel factors. In this approach, for each speaker, the speaker factor is assumed to be the same for all recordings of that particular class, while the channel factors are different for each recording.

In JFA [88], a given speaker GMM supervector M is assumed to be decomposable in this form:

$$M = \mu + Vy + Ux + Dz \tag{C.1}$$

where μ is a speaker- and session-independent super-vector (from UBM), V and D are the eigenvoice matrix and diagonal residual, respectively, and they define a speaker subspace. U defines a session subspace, the eigenchannel matrix. The vectors y and z are speaker-dependent factors, and vector x is a channel-dependent factor. All three vectors are assumed to be random variables with a normal distribution $N(0, I)$. The step-by-step procedure for applying JFA to speaker-recognition is as follows:

First train the eigenvoice matrix V, assuming that U and D are zero.

Second train the eigenchannel matrix U, given estimates of V, assuming that D is zero.

Third train residual matrix D, given estimates of V and U.

Forth using the computed matrices from previous steps, compute the speaker-dependent (y and z) and channel-dependent (x) factors.

Fifth calculate the score for test conversation side and target speaker conversation side, using the matrices and controlling factors [177].

This method of modelling, which separates subspaces for capturing speaker, channel, and session variabilities, proved [130] to suffer from miss-modelling of some speaker-specific information during training of the channel subspace.

"The cure for pain is in the pain."
Jalāl ad-Dīn Muhammad Rūmī

D
Appendix

D.1 Marginals and Conditionals of Gaussians

Prior to addressing factor analysis, it is necessary to explain the determination of conditional and marginal distributions of random variables of Gaussians. Assuming a vector-valued random variable is used,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{D.1}$$

where $x_1 \in R^r$, $x_2 \in R^s$, and $x \in R^{r+s}$. Suppose $x \sim \mathcal{N}(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \tag{D.2}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{D.3}$$

Here, $\mu_1 \in R^r$, $\mu_2 \in R^s$, $\Sigma_{11} \in R^{r\times r}$, and $\Sigma_{12} \in R^{r\times s}$. Furthermore $\Sigma_{12} = \Sigma_{21}^T$, as covariance matrices are symmetric.

Based on the working assumptions [178], it can be inferred that x_1 and x_2 are jointly multivariate Gaussian. To determine the marginal distribution of x_1, one must note that $E[x_1] = \mu_1$ and $Cov(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$. In order to certify the validity of the latter relationship and according to the joint covariance of x_1 and x_2, the following is obtained [178]

$$\begin{aligned} Cov(x) &= \Sigma \\ &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\ &= E[(x-\mu)(x-\mu)^T] \\ &= E\begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1)^T & (x_1-\mu_1)(x_2-\mu_2)^T \\ (x_2-\mu_2)(x_1-\mu_1)^T & (x_2-\mu_2)(x_2-\mu_2)^T \end{bmatrix} \end{aligned} \tag{D.4}$$

The result is attained from the association of the upper-left sub-blocks in the matrices in the second and final lines.

On the basis of the fact that the marginal distributions of Gaussians are Gaussian as well, the marginal distribution of x_1 is established to be denoted by $x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ [178].

Additionally, in accordance with the description of the multivariate Gaussian distribution [178], the conditional distribution of x_1 given x_2 can be determined to be $x1|x2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$, where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \tag{D.5}$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{D.6}$$

The above equations for the determination of the conditional and marginal distributions of Gaussians are of great use in the factor analysis model.

"One thing I have learned in a long life: that all our science, measured against reality, is primitive and childlike - and yet it is the most precious thing we have."

Albert Einstein

E Appendix

E.1 Implementation Steps for Calculating i-vectors from Speech Signal

Suppose we have a sequence of L frames $\{y_1, y_2, ..., y_L\}$ and an UBM composed of C mixture components defined in the feature space of dimension F. Initially, 0^{th}, $1^{st} and 2^{nd}$ order sufficient statistics need to be calculated for each of the training utterances (u), as follows,

$$N_c(u) = \sum_{t=1}^{L} \gamma_t(c) \tag{E.1}$$

$$F_c(u) = \sum_{t=1}^{L} \gamma_t(c) y_t \tag{E.2}$$

$$S_c(u) = diag\{\sum_{t=1}^{L} \gamma_t(c) y_t y_t^*\} \tag{E.3}$$

which $\gamma_t(c)$ is the posterior of gaussian component c for observation t of utterance u.

The second step is to compute the centralized $1^{st} and 2^{nd}$ order statistics,

$$\tilde{F}_c(u) = F_c(u) - N_c(u) mc \tag{E.4}$$

$$\tilde{S}_c(u) = S_c(u) - diag\{F_c(u)m_c^* + m_cF_c(u)^* - N_c(u)m_cm_c^*\} \tag{E.5}$$

where m_c is the UBM mean for mixture component c. By expanding the statistics into matrices we get,

$$NN(u) = \begin{pmatrix} N_1(u) * I & & \\ & \ddots & \\ & & N_c(u) * I \end{pmatrix} \tag{E.6}$$

$$FF(u) = \begin{pmatrix} \tilde{F}_1(u) \\ \vdots \\ \tilde{F}_c(u) \end{pmatrix} \tag{E.7}$$

$$SS(u) = \begin{pmatrix} \tilde{S}_1(u) & & \\ & \ddots & \\ & & \tilde{S}_c(u) \end{pmatrix} \tag{E.8}$$

Where I is the $F \times F$ identity matrix and $NN(u)$ and $SS(u)$ are the $CF \times CF$ block diagonal matrix and $FF(u)$ is the $CF \times 1$ vector obtained by concatenating $\tilde{F}_1(u), ..., \tilde{F}_c(u)$.

Third step is the initial estimation of the total factors w. Assume that $l(s)$ is the $R \times R$ matrix defined by:

$$l_T(u) = I + T^*\Sigma^{-1}NN(u)T \tag{E.9}$$

Where Σ^{-1} is the inverse of UBM covariance matrix. And R is the rank of total variability matrix, which needs to be less than $C \times F$. Proposition 1 in [89] (which was proposed for training of eigenvoice matrix), could be applied for T-matrix training as follows: for each utterance u, the posterior distribution of $w(u)$ given $x(u)$ and parameter set (random initialization of T, Σ from UBM) is Gaussian with $w(u) \sim Normal(l_T^{-1}(u)T^*\Sigma^{-1}FF(u), l_T^{-1}(u))$. The $E[w(u)]$ denotes the posterior expectation of $w(u)$, which is

$$E[w(u)] = l_T^{-1}(u)T^*\Sigma^{-1}FF(u) \tag{E.10}$$

To give a intuitive view on equation E.10, it could be proved that $E[w(u)]$ is a solution to the least-square's quadratic minimization problem of : $\min_{w(u)} ||FF(u) - Tw(u)||^2$

Forth step is to use current estimate of posterior distribution parameters of $w(u)$, for computing some additional statistics across utterances, these statistics are,

$$N_c = \sum_u N_c(u) \tag{E.11}$$

$$A_c = \sum_u N_c(u)l_T^{-1}(u) \tag{E.12}$$

$$\mathbb{C} = \sum_u FF(u) * (l_T^{-1}(u) * T^* * \Sigma^{-1} * FF(u))^* \tag{E.13}$$

$$NN = \sum_u NN(u) \tag{E.14}$$

In which $(l_T^{-1}(u) * T^* * \Sigma^{-1} * FF(u))$ and $l_T^{-1}(u)$ are mean and covariance of posterior distribution of w(u), respectively.

In the fifth step the T matrix estimated using accumulated statistics from previous step, as follows:

$$\begin{bmatrix} T_1 \\ \vdots \\ T_c \end{bmatrix} = \begin{bmatrix} A_1^{-1} * \mathbb{C}_1 \\ \vdots \\ A_c^{-1} * \mathbb{C}_c \end{bmatrix} \tag{E.15}$$

where $\mathbb{C} = \begin{bmatrix} \mathbb{C}_1 \\ \vdots \\ \mathbb{C}_c \end{bmatrix}$ is the block matrix components of c corresponding to each Gaussian mixture.

Finally, the last step is to run N (practically, N is between 10 and 20) iterations of 3^{rd} to 5^{th} steps, and substitute estimates of the T-matrix into equations in step 3.

Extracting the i-vector from training and testing utterances is the next step of front-end factor analysis. The method for extracting subspace factors has already been explained during the description of the steps of T-matrix training. To be specific, i-vectors are the posterior expectation of the distribution of subspace factors, which are assumed to be Gaussian. Given the T-matrix and sufficient statistics of utterance, u i-vectors are obtained by,

$$w(u) = (I + T^*\Sigma^{-1}NN(u)T)^{-1}.T^*\Sigma^{-1}FF(u) \tag{E.16}$$

Which is achieved by substitution of the equation E.9 into the equation E.10.

"Pure mathematics is, in its way, the poetry of logical ideas."

Albert Einstein

F
Appendix

F.1 Factor Analysis Vs. Principal Component Analysis

In Section 2.4.2 the factor analysis approach is discussed, and it gives a way to model data ($x \in R^n$) as approximately lying in some y-dimension subspace (Y-space), where $y << n$ (n is the dimension of original space, super-vector space). In this method it is imagined that each point $x^{(i)}$ was created by generating some $w^{(i)}$ in y-dimension affine space $\{M = \mu + Tw; w \in R^y\}$, and then adding Υ-covariance noise. However, PCA will tackle the problem more directly, and will require only an eigenvector calculation, and the classical PCA does not need to resort to EM.

Assume we have the following dataset (crosses on the Figure F.1.1). Now, suppose we pick u to correspond the the direction shown in the Figure F.1.1. The circles denote the projections of the original data onto this line. We see that the projected data still has a fairly large variance, and the points tend to be far from zero. In contrast, suppose had instead picked another direction, r, Here, the projections have a significantly smaller variance, and are much closer to the origin. PCA automatically select the direction u corresponding to the u shown above.

Figure F.1.2 shows the typical samples of $w^{(i)}$ in a one-dimensional sub-space. Then these data-points are mapped to the two dimensional s space, by $\mu + Tw$. This model envisioning that X's inside each mono Gaussian circles are considered as original data points $x^{(i)}$. The working assumption is that the sampling of a y dimension mono Gaussian $w^{(i)}$ produces every data point $x^{(i)}$. The subsequent calculation $\mu + Tw^{(i)}$ enables the mapping of $x^{(i)}$ to a y-dimensional affine subspace of R^s. Afterwards, the addition of covariance Υ noise to $\mu + Tw^{(i)}$ generates $x^{(i)}$.

Consider the factor analysis model where we constrain $\Upsilon = \sigma^2 I$, and T to be orthonormal. It

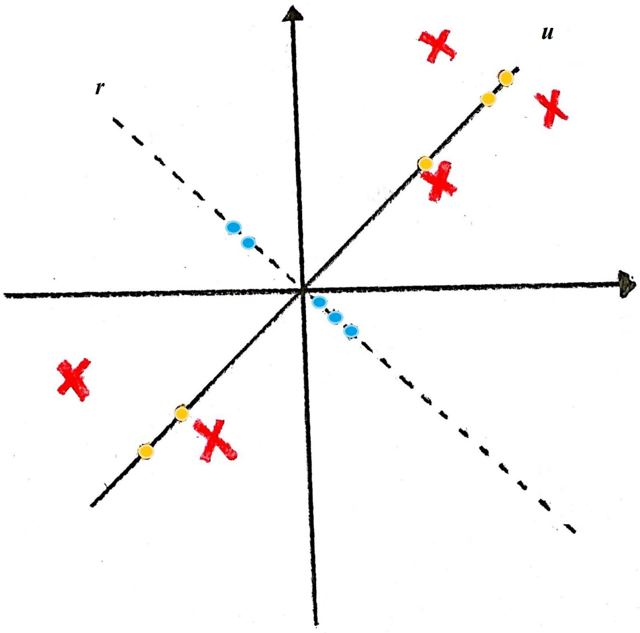


Figure F.1.1: Graphical example of classical PCA approach.

is shown that [179], as $\sigma^w- > 0$ this model reduces to classical PCA, also known as Karhunen Loeve transform.

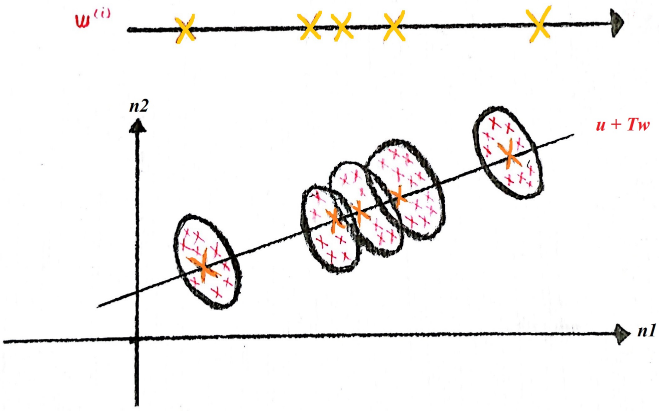


Figure F.1.2: Graphical example of factor analysis based approach.

References

[1] A. Hanani, *Human and computer recognition of regional accents and ethic groups from british english speech*. PhD thesis, School of Electronic, Electrical and Computer Engineering, 2012.

[2] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[3] N. Dehak and S. Shum, "Low-dimensional speech representation based on factor analysis and its applications," in *Interspeech*, 2011.

[4] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. J. Russell, S. Steidl, and M. Wong, "The pf_star children's speech corpus.," in *INTERSPEECH*, pp. 2761–2764, 2005.

[5] R. Auckenthaler and J. Mason, "Uws submission to nist 2003," tech. rep., University Of Wales Swansea, 2003.

[6] R. Auckenthaler, *Texy-Independent Speaker Verification with Limited Resources*. PhD thesis, Department of Electrical and Electronic Engineerin, Univerdity of Wales Swansea., May 2001.

[7] S. Ghai, *Addressing pitch mismatch for children's automatic speech recognition*. PhD thesis, Department of electronics and electrical engineering, Indian institute of technology Guwahat, 2011.

[8] D. A. Reynolds and L. P. Heck, "Automatic speaker recognition," in *AAAS 2000 Meeting, Humans, Computers and Speech Symposium*, vol. 19, pp. 101–104, 2000.

[9] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, pp. 65–78, Feb 2002.

[10] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[11] A. Potamianos and S. Narayanan, “Robust recognition of children’s speech,” *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 603–616, Nov 2003.

[12] S. Lee, A. Potamianos, and S. Narayanan, “Analysis of children’s speech: duration, pitch and formants.,” in *EUROSPEECH*, ISCA, 1997.

[13] S. Schötz, “A perceptual study of speaker age,” *Lund Working Papers in Linguistics*, vol. 49, pp. 136–139, 2009.

[14] M. Iseli, Y.-L. Shue, and A. Alwan, “Age-and gender-dependent analysis of voice source characteristics,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006.

[15] B. Weinrich, B. Salz, and M. Hughes, “Aerodynamic measurements: Normative data for children ages 6:0 to 10:11 years,” *Journal of Voice*, vol. 19, no. 3, pp. 326 – 339, 2005.

[16] A. Potamianos and S. Narayanan, “A review of the acoustic and linguistic properties of children’s speech,” in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pp. 22–25, Oct 2007.

[17] M. Gerosa, D. Giuliani, and F. Brugnara, “Acoustic variability and automatic recognition of children’s speech,” *Speech Communication*, vol. 49, no. 11, pp. 847 – 860, 2007. Intrinsic Speech Variations.

[18] S. Eguchi and I. J. Hirsh, “Development of speech sounds in children.,” *Acta otolaryngologica. Supplementum*, vol. 257, pp. 1–51, 1969.

[19] R. D. Kent, “Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies,” *Journal of Speech and Hearing Research*, vol. 19, no. 3, p. 421, 1976.

[20] Q. Li and M. J. Russell, “Why is automatic recognition of children’s speech difficult?,” in *INTERSPEECH* (P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, eds.), pp. 2671–2674, ISCA, 2001.

[21] M. Benzeguiba, R. De-Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, “Automatic speech recognition and intrinsic speech variation,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5, May 2006.

[22] R. D. Kent and L. Forner, “Speech segment duration in sentence recitations by children and adults.,” *Journal of Phonetics*, p. 157⊠168, 1980.

[23] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech an acoustic study of consonants and consonant-vowel transition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–393–I–396, IEEE, 2006.

[24] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[25] A. S.-L.-H. Association, "typical speech and language development," 1997-2015.

[26] E. Strommen and F. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, no. 1, pp. 5–16, 1993.

[27] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, pp. 197–200 vol.1, May 1998.

[28] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, pp. 1455–1468, 1999.

[29] Q. Li and M. J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Seventh International Conference on Spoken Language Processing*, 2002.

[30] S. Yildirim, S. Narayanan, D. Boyd, and S. Khurana, "Acoustic analysis of preschool children's speech," in *Proc. 15th ICPhS*, p. 949–952, 2003.

[31] L. Besacier, J. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2-3, pp. 89–106, 2000.

[32] S. Ullah and F. Karray, "Hybrid feature selection approach for natural language call routing systems," in *Information and Emerging Technologies, 2007. ICIET 2007. International Conference on*, pp. 1–5, July 2007.

[33] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using gaussian mixture models," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, pp. 343–346, Dec 2001.

[34] M. Feld, E. Barnard, C. van Heerden, and C. Muller, "Multilingual speaker age recognition: Regression analyses on the lwazi corpus," in *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, pp. 534–539, Nov 2009.

[35] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1044–1046, 2007.

[36] S. Safavi, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification.," *IEEE Signal Processing Letters.*, vol. 19, no. 12, pp. 829–832, 2012.

[37] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Speaker recognition for children's speech," *Interspeech*, pp. 1836–1839, 2012.

[38] A. H. M. R. S. Safavi, M. Najafian and P. Jancovic, "Comparison of speaker verification performance for adult and child speech," in *WOCCI*, 2014.

[39] S. Safavi, P. Jančovič, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans.," in *INTERSPEECH*, pp. 2440–2444, ISCA, 2013.

[40] S. Safavi, M. J. Russell, and P. Jančovič, "Identification of age-group from children's speech by computers and humans.," in *INTERSPEECH*, pp. 2440–2444, ISCA, 2013.

[41] D. Boatman, "The neurocognition of language.," *Brain*, vol. 125, no. 1, pp. 215–216, 2002.

[42] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: Results of functional data analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[43] E. MacDonald, E. Johnson, J. Forsythe, P. Plante, and K. Munhall, "Children's development of self-regulation in speech production," *Current Biology*, vol. 22, no. 2, pp. 113 – 117, 2012.

[44] K. J. Ballard, D. A. Robin, G. Woodworth, and L. D. Zimba, "Age-related changes in motor control during articulator visuomotor tracking," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 4, pp. 763–777, 2001.

[45] F. G. DiSimoni, "Influence of consonant environment on duration of vowels in the speech of three, six, and nine year old children," *The Journal of the Acoustical Society of America*, vol. 55, no. 2, pp. 362–363, 1974.

[46] R. D. Kent and L. Forner, "Speech segment duration in sentence recitations by children and adults.," *Journal of Phonetics*, pp. pp.157–168, 1980.

[47] S. G. Sharkey and J. W. Folkins, "Variability of lip and jaw movements in children and adultsimplications for the development of speech motor control," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 8–15, 1985.

[48] W. D. Voiers, "Perceptual bases of speaker identity," *The Journal of the Acoustical Society of America*, vol. 36, p. 1065, 1964.

[49] G. L. Holmgren, "Physical and psychological correlates of speaker recognition," *Journal of Speech, Language and Hearing Research*, vol. 10, no. 1, p. 57, 1967.

[50] F. R. Clarke and R. W. Becker, "Comparison of techniques for discriminating among talkers," *Journal of Speech, Language and Hearing Research*, vol. 12, no. 4, p. 747, 1969.

[51] R. N. A. D. H. Perry, Theodore L.; Ohde, "The acoustic bases for gender identification from children's voices," *The Journal of the Acoustical Society of America*, vol. 109, pp. 2988–2998, 2001.

[52] C. Mathon and S. de Abreu, "Emotion from speakers to listeners: Perception and prosodic characterization of affective speech," in *Speaker Classification II* (C. Muller, ed.), vol. 4441 of *Lecture Notes in Computer Science*, pp. 70–82, Springer Berlin Heidelberg, 2007.

[53] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[54] L. Rabiner and R. Schafer, *Digital processing of speech signals*. Prentice-Hall signal processing series, Prentice-Hall, 1978.

[55] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[56] J. Tierney, "A study of lpc analysis of speech in additive noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 389–397, 1980.

[57] H. Hermansky and J. Cox, L.A., "Perceptual linear predictive (plp) analysis-resynthesis technique," in *Applications of Signal Processing to Audio and Acoustics, 1991. Final Program and Paper Summaries., 1991 IEEE ASSP Workshop on*, pp. 37–38, Oct 1991.

[58] M. Hossan, S. Memon, and M. Gregory, "A novel approach for mfcc feature extraction," in *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, pp. 1–5, Dec 2010.

[59] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, pp. IV–788–91 vol.4, 2003.

[60] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: report from jhu ws'02," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, pp. IV–792–5 vol.4, 2003.

[61] W. D. Andrews, M. A. Kohler, J. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–149–I–152, 2002.

[62] J. Navratil, Q. Jin, W. D. Andrews, and J. Campbell, "Phonetic speaker recognition using maximum-likelihood binary-decision tree models," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, pp. IV–796–9 vol.4, 2003.

[63] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. D. Andrews, and J. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, pp. IV–800–3 vol.4, 2003.

[64] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell, "Conditional pronunciation modeling in speaker detection," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 4, pp. IV–804–7 vol.4, 2003.

[65] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers.," in *INTERSPEECH* (P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, eds.), pp. 2521–2524, ISCA, 2001.

[66] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "Sri's 2004 nist speaker recognition evaluation system," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, pp. 173–176, 2005.

[67] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[68] M. Sambur, "Selection of acoustic features for speaker identification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 176–182, 1975.

[69] C. S. Gupta, "Significance of source features for speaker recognition," Master's thesis, Department of Computer Science and Engineering, Indiin Institute of Technology., April 2003.

[70] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features.," in *INTERSPEECH*, 2002.

[71] D. Naik, "Pole-filtered cepstral mean subtraction," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, pp. 157–160 vol.1, May 1995.

[72] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 578–589, Oct 1994.

[73] A. Garcia and R. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 325–328 vol.1, 1999.

[74] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133 – 147, 1998.

[75] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, pp. 1–4, 2005.

[76] F. Allen, E. Ambikairajah, and J. Epps, "Warped magnitude and phase-based features for language identification," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, 2006.

[77] T. Dietterich, *Advances in neural information processing systems 14. 1*. No. v. 14, MIT Press, 2002.

[78] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, Sep 1997.

[79] H. Gish and M. Schmidt, "Text-independent speaker identification," *Signal Processing Magazine, IEEE*, vol. 11, pp. 18–32, Oct 1994.

[80] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[81] J. A. Bilmes *et al.*, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[82] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 291–298, Apr 1994.

[83] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, 2006.

[84] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[85] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[86] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, pp. 308–311, May 2006.

[87] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[88] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[89] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[90] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[91] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4237–4240, 2009.

[92] Odyssey, *The MITLL NIST LRE 2011 Language Recognition System*, Odyssey, June 2012.

[93] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 13–16 vol.1, Mar 1992.

[94] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation approaches for speaker verification," *Information Forensics and Security, IEEE Transactions on*, vol. 5, pp. 802–809, Dec 2010.

[95] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006 The*, pp. 1–6, IEEE, 2006.

[96] M. Przybocki, A. Martin, and A. Le, "Nist speaker recognition evaluation chronicles - part 2," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pp. 1–6, June 2006.

[97] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," tech. rep., DTIC Document, 1997.

[98] A. F. Martin and M. A. Przybocki, "The nist speaker recognition evaluations: 1996-2001," in *Proc. of SPIE Vol*, vol. 7324, pp. 732411–1, 2001.

[99] V. Wan and W. Campbell, "Suppoer vector machines for speaker verification and identification," in *Proceedings of the 2000 IEEE signal processing society workshop.*, vol. 2, pp. 775–784, 2000.

[100] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19 – 41, 2000.

[101] R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[102] W. Campbell and Z. N. Karam, "A framework for discriminative svmgmm systems for language recognition," in *Tenth Annual Conference of the International Speech Communication Association*, pp. 2195–2198, 2009.

[103] R. Leonard, G. Doddington, and T. I. I. DALLAS., *Automatic Language Identification*. Defense Technical Information Center, 1974.

[104] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. i. preliminary methodological considerations," *The Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 708–713, 1977.

[105] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech Language*, vol. 27, no. 1, pp. 59–74, 2013.

[106] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, pp. 31–, Jan 1996.

[107] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors.," in *INTERSPEECH*, pp. 745–748, ISCA, 2011.

[108] A. DeMarco and S. J. Cox, "Iterative classification of regional british accents in i-vector space," in *Symposium on Machine Learning in Speech and Language Processing (SIGML)*, 2012.

[109] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Dimension reduction approaches for svm based speaker age estimation.," in *INTERSPEECH*, pp. 2031–2034, ISCA, 2009.

[110] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge.," in *INTERSPEECH*, pp. 2794–2797, 2010.

[111] M. Kockmann, L. Burget, and J. Cernocky, "Brno university of technology system for interspeech 2010 paralinguistic challenge.," in *INTERSPEECH* (T. Kobayashi, K. Hirose, and S. Nakamura, eds.), pp. 2822–2825, ISCA, 2010.

[112] D. Mahmoodi, A. Soleimani, H. Marvi, F. Razzazi, M. Taghizadeh, and M. Mahmoodi, "Age estimation based on speech features and support vector machine," in *Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd*, pp. 60–64, July 2011.

[113] C.-C. Chen, P.-T. Lu, M.-L. Hsia, J.-Y. Ke, and O.-C. Chen, "Gender-to-age hierarchical recognition for speech," in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, pp. 1–4, Aug 2011.

[114] C. van Heerden, E. Barnard, M. Davel, C. van der Walt, E. van Dyk, M. Feld, and C. Muller, "Combining regression and classification methods for improving automatic speaker age recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5174–5177, March 2010.

[115] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1605–1608, 2008.

[116] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via I-Vectors and Dimensionality Reduction," in *INTERSPEECH 2011*, (Florence, Italy), pp. 857–860, Aug. 2011.

[117] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4 – 39, 2013. Special issue on Paralinguistics in Naturalistic Speech and Language.

[118] T. Bocklet, G. Stemmer, V. Zeissler, and E. Noth, "Age and gender recognition based on multiple systems - early vs. late fusion," *Interspeech*, pp. 2830–2833, 2010.

[119] R. Porat, D. Lange, and Y. Zigel, "Age Recognition Based on Speech Signals Using Weights Supervector," in *INTERSPEECH*, pp. 2814–2817, 2010.

[120] P. Nguyen, T. Le, D. Tran, X. Huang, and D. Sharma, "Fuzzy Support Vector Machines for Age and Gender Classification," in *INTERSPEECH*, pp. 2806–2809, 2010.

[121] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE*, vol. 3, pp. 72–83, 1995.

[122] J. Naik, L. Netsch, and G. Doddington, "Speaker verification over long distance telephone lines," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pp. 524–527 vol.1, May 1989.

[123] M. F. BenZeghiba and H. Bourlard, "User-customized password speaker verification using multiple reference and background models," *Speech Communication*, vol. 48, no. 9, pp. 1200 – 1213, 2006.

[124] L. P. Heck, Y. Konig, M. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2-3, pp. 181 – 192, 2000.

[125] B. Yegnanarayana and S. Kishore, "Aann: an alternative to {GMM} for pattern recognition," *Neural Networks*, vol. 15, no. 3, pp. 459 – 469, 2002.

[126] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 210 – 229, 2006. Odyssey 2004: The speaker and Language Recognition Workshop Odyssey-04 Odyssey 2004: The speaker and Language Recognition Workshop.

[127] H. Ezzaidi and J. Rouat, "Pitch and mfcc dependent gmm models for speaker identification systems," in *Electrical and Computer Engineering, 2004. Canadian Conference on*, vol. 1, pp. 43–46 Vol.1, May 2004.

[128] M. Newman, L. Gillick, Y. Ito, D. Mcallaster, and B. Peskin, "Speaker verification through large vocabulary continuous speech recognition," in *Proc. ICSLP*, pp. 2419–2422, 1996.

[129] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *Signal Processing Magazine, IEEE*, vol. 26, pp. 95–103, March 2009.

[130] N. Dehak, *Discriminative and Generative Approaches for Long- and Short-term Speaker Characteristics Modeling: Application to Speaker Verification.* PhD thesis, 2009. AAINR50490.

[131] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.," in *INTERSPEECH*, vol. 9, pp. 1559–1562, 2009.

[132] G. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar, "S.: Duration and pronunciation conditioned lexical modeling for speaker verification," in *In: Proceedings of Interspeech*, 2007.

[133] M. Bahari and H. Van Hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2011 IEEE Workshop on*, pp. 1–6, Sept 2011.

[134] T. Bocklet, A. Maier, and E. Noth, "Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines regression.," in *TSD*, vol. 5246 of *Lecture Notes in Computer Science*, pp. 253–260, Springer, 2008.

[135] D. A. van Leeuwen and M. H. Bahari, "Calibration of probabilistic age recognition.," in *INTERSPEECH*, ISCA, 2012.

[136] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1975–1985, Sept 2011.

[137] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–1089–IV–1092, April 2007.

[138] M. H. Bahari, M. McLaren, H. V. Hamme, and D. A. van Leeuwen, "Age estimation from telephone speech using i-vectors.," in *INTERSPEECH*, ISCA, 2012.

[139] A. D. Rogol, P. A. Clark, and J. N. Roemmich, "Growth and pubertal development in children and adolescents: effects of diet and physical activity," *The American journal of clinical nutrition*, vol. 72, no. 2, pp. 521s–528s, 2000.

[140] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech Language*, vol. 27, no. 1, pp. 151–167, 2013.

[141] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," *Int. Conf. on Spoken Language Processing*, 2000.

[142] N. M. I. Group., "2003 nist speaker recognition evaluation ldc2010s03.," in *Philadelphia: Linguistic Data Consortium*, 2010.

[143] R. Bolle, S. Pankanti, and N. Ratha, "Evaluation techniques for biometrics-based authentication systems (frr)," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, pp. 831–837 vol.2, 2000.

[144] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[145] S. M. DArcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson, "The accents of the british isles (abi) corpus," *Proceedings Modelisations pour l'Identification des Langues*, pp. 115–119, 2004.

[146] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[147] S.-H. Chen and Y.-R. Luo, "Speaker verification using mfcc and support vector machine," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, IMECS, March 2009.

[148] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, pp. 52–59, Feb 1986.

[149] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pp. 532–535, IEEE, 1989.

[150] M. S. Seyed Omid Sadjadi and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter, IEEE*, 2013.

[151] P. A. Torres-Carrasquillo, E. Singer, W. M. Campbell, T. P. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, and D. E. Sturim, "The mitll nist lre 2007 language recognition system.," in *INTERSPEECH*, pp. 719–722, ISCA, 2008.

[152] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 293–296 vol. 1, IEEE, 1990.

[153] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[154] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrataz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

[155] D. R. Miller and J. Trischitta, "Statistical dialect classification based on mean phonetic features," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4, pp. 2025–2027 vol. 4, IEEE, 1996.

[156] L. M. Arslan and J. H. L. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[157] C. Teixeira, I. Trancoso, and A. Serralheiro, "Recognition of non-native accents," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[158] M. Lincoln, S. Cox, and S. Ringland, "A comparison of two unsupervised approaches to accent identification," 1998.

[159] R. Huang, J. Hansen, and P. Angkititrakul, "Dialect/accent classification using unrestricted audio," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, pp. 453–464, 2007.

[160] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 634–646, 2006.

[161] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 1, p. 31, 1996.

[162] J. C. Wells, *Accents of English*, vol. 1. Cambridge University Press, 1982.

[163] S. Nittrouer and D. Whalen, "The perceptual effects of child and adult differences in fricative-vowel coarticulation," *J. Acoust. Soc. Am.*, vol. 86, pp. 1266–1276, 1989.

[164] T. Roberts and C. A. Will, "Adaptive speaker verification apparatus and method including alternative access control," Sept. 12 2000. US Patent 6,119,084.

[165] K. Prasad, P. Lotia, and M. Khan, "A review on text-independent speaker identification using gaussian supervector svm," *International Journal of u-and e-Service, Science and Technology*, vol. 5, no. 1, pp. 71–82, 2012.

[166] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *Signal Processing Letters, IEEE*, vol. 2, no. 3, pp. 46–48, 1995.

[167] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, no. 4, pp. 312–322, 2008.

[168] E. S. Parris and M. J. Carey, "Discriminative phonemes for speaker identification," in *Third International Conference on Spoken Language Processing*, 1994.

[169] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[170] K. Wu and D. G. Childers, "Gender recognition from speech. Part i: Coarse analysis," *Journal of the Acoust. Soc. Am.*, vol. 90, no. 4, pp. 1828–1840, 1991.

[171] E. S. Parris and M. J. Carey, "Language independent gender identification," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, no. 685-688, 1996.

[172] C. Müuller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," *Interspeech*, pp. 2118–2121, 2006.

[173] P. Jančovič and M. Köküer, "Estimation of voicing-character of speech spectra based on spectral shape," *IEEE Signal Processing Letters*, vol. 14, pp. 66–69, Jan. 2007.

[174] M. Li, C.-S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," *Interspeech*, 2010.

[175] K. Wu and D. G. Childers, "Gender recognition from speech. Part ii: Fine analysis," *Journal of the Acoust. Soc. Am.*, vol. 90, no. 4, pp. 1841–1856, 1991.

[176] R. V. Hogg and J. Ledolter, *Engineering statistics*. Macmillan Pub Co, 1987.

[177] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4057–4060, IEEE, 2009.

[178] A. Ng., "Factor analysis," tech. rep., Stanford, 2011.

[179] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.