Automatic Speaker Classification Based on Voice Characteristics

A thesis submitted for the degree
of Master of Information Sciences (Research) of
the University of Canberra

Phuoc Thanh Nguyen

December 2010

Summary of Thesis

Gender, age, accent and emotion are some of speaker characteristics being investigated in voice-based speaker classification systems. Classifying speaker characteristics is an important task in the fields of Dialog, Speech Synthesis, Forensics, Language Learning, Assessment, and Speaker Recognition.

It is well known that reducing classification error rate has been a challenge in those research fields. This research thesis investigates new methods for speech feature extraction and classification to meet this challenge. Extracted speech features range from traditional features in speech recognition such as mel-frequency cepstral coefficients (MFCCs) to recently developed prosodic and voice quality features in speaker classification such as pitch, shimmer and jitter. Feature selection was then performed to find a more suitable feature set for building speaker models. For classification methods, feature weighting vector quantisation, Gaussian mixture models (GMMs), Support Vector Machine (SVM) and Fuzzy Support Vector Machine (FSVM) are investigated. Those new feature extraction and classification methods are then applied to gender, age, accent and emotion classification. Four well-known data sets including Australian National Database of Spoken Language (ANDOSL), aGender, EBO-DB, and FAU AIBO are used to evaluate those methods.

The contributions of this thesis to classification of speaker characteristics include:

1. The use of different speech features. Up to 1582 features and transliteration have been investigated.

2. Application of new feature selection method. Correlation based feature subset selection with SFFS was employed to eliminate redundant features because of large databases.

3. The use of fuzzy SVM (FSVM) as a new speaker classification method. FSVM assigns

a fuzzy membership value as a weight to each training data point to allow the decision boundary to move to overlapping regions to reduce empirical errors.

4. A detailed comparison of speaker classification performance for GMMs, SVM and FSVM.

5. A depth investigation on the relevance of feature type for classification of age and gender. Extensive experiments are performed to determine which features in the speech signal are suited to representation of age and gender in human speech.

6. Classification of age, gender, accent, and emotion characteristics is performed on four well-known data sets including ANDOSL, aGender, EBO-DB and FAU AIBO.

Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled —

Automatic Speaker Classification Based on Voice Characteristics

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in *Gold Book Part 7: Examination of Higher Degree by Research Theses Policy, Schedule Two (S2).*
Refer to http://www.canberra.edu.au/research-students/goldbook

Signature of Candidate

Signature of chair of the supervisory panel

Date

Acknowledgements

First and foremost, I would like to thank my supervisor, A/Prof. Dat Tran, for his enormous support during my study at the University of Canberra. I am also thankful for his valuable guidance both in research and life, his encouragement and attention to important research milestones and events, his very quick response to my questions, and his patience to help me enhance the thesis.

I would also like to thank my co-supervisor, A/Prof. Xu Huang, for his encouragement, advice, support and suggestions on research plans. I am also thankful for his patience to revised my thesis and careful feedbacks.

I would also like to thank the Faculty of Information Sciences and Engineering for supporting conference travels and maintaining the excellent computing facilities which were crucial for carrying out my research. Thanks to staff members as well as research students for discussions and seminars. A grateful thanks to Prof. John Campbell for his Research Proposal and Research Methodologies courses. A warm thanks to Mr. Hanh Huynh for his interesting discussions about life and encouragement. A special thanks to Trung Le for his valuable discussions.

More importantly, I would like to thank the HCMC University of Pedagogy, Viet Nam for providing me the scholarship which enabled me to undertake this research at the University of Canberra. I would like to express my gratitude to all my lecturers and colleagues at the Faculty of Mathematics and Informatics, HCMC University of Pedagogy. I wish to express my warm and sincere thanks to Dr. Nguyen Thai Son and Msc. Ly Anh Tuan, Faculty of Mathematics and Informatics, HCMC University of Pedagogy for their important guidance, support and encouragement during my first steps in the Faculty.

I devote my deepest gratitude to my parents for their unlimited love and support. They have encouraged me throughout the years of my study. The most special thanks belong to my wife Huyen, for her understanding about my leaving during all these years of my absence, her selfless love and support all along and encouragement.

Contents

List of Figures

List of Tables

Abbreviation

GMMs	Gaussian Mixture Models
SVM	Support Vector Machine
FSVM	Fuzzy Support Vector Machine
HTK	Hidden Markov Model Toolkit
HMM	Hidden Markov Model
SFFS	Sequential Forward Floating Search
MFCCs	Mel-Frequency Cepstral Coefficients
LPC	Linear Prediction Coding

Chapter 1

Introduction

1.1 Speaker Characteristics and Their Applications

Humans are very good at recognizing people. They can guess a person's gender, age, accent, and emotion by just hearing the person's voice over the phone. At the highest level, people use semantics, diction, idiolect, pronunciation and idiosyncrasies, which emerge from socio-economic status, education and place of birth of a speaker. At the intermediate level, they use prosodic, rhythm, speed, intonation and volume of modulation, which discriminate personality and parental influence of a speaker. At the lowest level they use acoustic aspects of sounds, such as nasality, breathiness or roughness [56]. Recordings of the same utterance of two people will sound different because the process of speaking engages the individual mental and physical systems. Since these systems are different among people, their speech will be also different even for the same message. The speaker-specific characteristics in the signal can be exploited by listeners and technological applications to describe and classify speakers, based on age, gender, accent, language, emotion or health [16].

There are many speaker characteristics that have useful applications. The most popular of these include gender, age, health, language, dialect, accent, sociolect, idiolect, emotional state and attentional state [56]. These characteristics have many applications in Dialog Systems, Speech Synthesis, Forensics, Call Routing, Speech Translation, Language Learning, Assessment Systems, Speaker Recognition, Meet-

ing Browser, Law Enforcement, Human-Robot Interaction, and Smart Workspaces. For example, the Spoken Dialogs Systems provide services in the domains of finance, travel, scheduling, tutoring or weather. The systems need to gather information from the user automatically in order to provide timely and relevant services. Most telephone-based services today use spoken dialog systems to either route calls to the appropriate agent or even handle the complete service by an automatic system [56].

Some of the reasons for automatic speaker classification include: automatic indexing of audio material, identification or verification of people to ensure secure access, loading pre-trained models for speech recognition tasks, tailoring machine dialogue to the needs and situation of the user, or synthesizing voice with similar characteristics (gender, age, accent) to the speaker [30]. Demand for human-like response systems is increasing. For example, shopping systems can recommend suitable goods appropriate to the age and sex of the shopper.

1.2 Gender, Age, Accent and Emotion Classification

Gender, age, accent, and emotion have received a lot of attention in the area of speaker classification because of the increasing applications set out above. There are open challenges for participants to build systems and try to increase the accuracy of these speaker classification tasks [51, 52].

Gender classification achieved high accuracy, 94% on NIST 1999 database of telephone speech [47], 95.4% on data collected from a deployed customer-care system, AT&T's "How May I Help You" system [59]. Most speaker classification systems differentiate gender at the first stage to improve their performance.

Every person goes through the process of ageing. Changes in our voices happen not only in early childhood and puberty but also in our adult lives into old age. A lot of acoustic features vary with speaker age. Acoustic variation has been found in temporal as well as in laryngeally and supralaryngeally conditioned aspects of speech [57]. Elderly people often speak slower than younger people; however, there is no

difference in articulation rate between young and old women during read speech [1]. It is found that the age of younger people often is overestimated, while the age of older people is underestimated [1]. This means the middle age range is usually longer than younger and older age range. Identifying age of elderly and non-elderly people is quite an easy task with high accuracy of 95% [39]. Usually the division into three or four age groups is used. Three age groups of young, middle age and elderly in ANDOSL corpus [38] were used. Four age groups of child, youth, adult and senior were used in aGender corpus for INTERPSPEECH 2010 paralinguistics challenge [52].

Accents can be confused with dialects. Accents are the variances of pronunciation of a language, while dialects are varieties of language differing in vocabulary, syntax, and morphology, as well as pronunciation. For example, British Received Pronunciation is an accent of English, while Scottish English is a dialect because it usually has grammatical differences, such as "Are ye no going?" for "Aren't you going?" [56]. Another example of accent is most British English accents differentiate the words Kahn, con and corn using three different back open vowel qualities; however many American English accents use only two vowels in the three words (e.g. Kahn and con become homophones) [56]. Speaker accent recognition has been applied in providing product ratings over cell-phones to consumers via a toll-free number [71]. The system only provides the necessary information by adapting to consumer profiles and eventually targeted advertising based on consumer demographics. Accents spoken by elderly speakers are usually heavier than younger speakers. As well, men tend to be more dialectal than women [30]. Accent is known to affect speech recognition performance a lot. This lead to the approach of accent-specific speech recognisers. Unfortunately this approach is challenged by the limited system resources and data. Particularly, embedded environments such as mobile or automotive applications limit the integration of multiple recognizers within one system [56].

Emotion recognition has found a lot of research interests recently [51]. The current emotion databases include acted (DES, EMO-DB), induced (ABC, eNTERFACE), and natural emotion (AVIC, SmartKom, SUSAS, VAM). Acted and induced emotions are also called prototypical emotions, and natural emotion is called spontaneous

emotion. The emotion spoken content can be predefined (DES, EMO-DB, SUSAS, eNTERFACE) or variant (ABC, AVIC, SAL, SmartKom, VAM) [4]. Emotions can be grouped into arousal (i.e. passive vs. active) and valence (i.e. positive vs. negative) in binary emotion classification tasks [53]. Spontaneous emotion data are harder to collect and label than prototypical emotion data. Emotion classification performances are higher on those prototypical databases than spontaneous ones. One way to increase performance of emotion classification is to employ speaker-dependent models. However the community's orientation is towards speaker independence because of its reality. Moreover, it is difficult to collect enough emotional data from an individual. The reason for the low speaker-independent classification performance is the differences of acoustic features between individual speakers, features can be multifunctional and inter-labeller agreement is - for spontaneous speech - not very high [4].

1.3 Research Problems

Emotions have various dimensional presentations and correlating these dimensions with acoustic features is difficult despite many approaches of division and experiments [19]. Researching emotion, however, is extremely challenging in several respects. One of the main difficulties results from the fact that it is difficult to define what emotion means in a precise way. There are ongoing debates concerning how many emotion categories exist, how to reconcile long-term properties such as moods with short-term emotional states such as full blown emotions, and debate as to how to seek measurable correlates of emotions. Hence, an engineering approach to emotion invariably has to rely on a number of assumptions to the problem for tractability [35].

At first glance, it may appear that we should be able to separate speaker characteristics from message characteristics in a speech signal quite easily. There is a view that speaker characteristics are predominantly low level - related to the implementation in a particular physical system of a given set of phonetic gestures, while message characteristics operate at a more abstract level - related to the choice of phonetic

gestures: the syllables, words and phrases that are used to communicate the meaning of a message. However this is to oversimplify the situation. Speakers are actually different at all levels, because speakers also differ in the way in which they realise the phonetic gestures, they vary in the inventory of gestures used, in the way in which gestures are modified by context, and in their frequency of use of gestures, words and message structure [16].

Children's speech is much more difficult than adult's speech in automatic speech recognition. This problem is even more difficult because of little training data. However, some approaches exist which try to compensate for this drawback. One remaining problem is the strong anatomic alteration of the vocal tract of children within a short period of time. An idea to solve this problem is to use different acoustic models for different age classes of children [6]. The most appropriate acoustic model has to be selected before the automatic speech recognition can be performed. If the age of a child is not known in advance, it can be predicted from the child's voice.

The INTERSPEECH 2009 emotion challenge and the INTERSPEECH 2010 paralinguistic challenge are two challenges for emotion, age and gender classification in the well-known INTERSPEECH conference. These challenges provide standardised corpora and test-conditions for participants to compare performances under exactly the same conditions in order to face more realistic scenarios of emotion, gender, age, and affect recognition [51, 52]. Accuracies for those characteristic classifications are still low, 38.2% in 5-class emotion classification [51], 81.2% in 3-class classification of male, female, and children, 48.9% in 4-class age classification [52]. Feature investigations and classification techniques have been conducted to increase accuracy. Some investigation on a good feature set for age and emotion have been worked out. These include acoustic, prosodic and linguistic features. However, there are still some questions. First, will a good feature set be different on different databases? Second, linguistic features will be different between databases because of different vocabulary. Third, a good feature set for classifying age, gender, and accent at the same time have not been studied. On the other hand, most studies are conducted on feature selection for speaker classification using popular classification techniques. There is little

research on a new classifier for classifying speaker characteristics. In this research, we make comparisons between GMMs and SVM performance and develop a Fuzzy support vector machine, an extension of Support vector machine, into speaker classification. Meanwhile, there has not been a system to classify speaker age, gender, and accent in one system. Additionally there has been no research on Australian accents. All these research question are included in my thesis.

Although the accent is only spoken by a minority of the population, it has a great deal of cultural credibility. It is disproportionately used in advertisements and by newsreaders. Current research on Australian accents and dialect focuses on the linguistic approach to dialect of phonetic study [5, 28], classification of native and non-native Australian [34], or to improve Australian automatic speech recognition performance [2]. However, there is no research on automatic speaker classification based on the three Australian accents of Broad, General, and Cultivated. According to linguists, three main varieties of spoken English in Australia are Broad (spoken by 34% of the population), General (55%) and Cultivated (11%) [40]. They are part of a continuum, reflecting variations in accent. Although some men use the accent, the majority of Australians that speak with the accent are women. Broad Australian English is usually spoken by men, probably because this accent is associated with Australian masculinity. It is used to identify Australian characters in non-Australian media programs and is familiar to English speakers. The majority of Australians speak with the General Australian accent. Cultivated Australian English has some similarities to British Received Pronunciation, and is often mistaken for it. In the past, the cultivated accent had the kind of cultural credibility that the broad accent has today. For example, until 30 years ago newsreaders on the government funded ABC had to speak with the cultivated accent [3].

1.4 Contributions of the Thesis

The research thesis presents the following contributions to classification of speaker characteristics:

1. The use of different voice features in speaker classification. Those voice features are as follows: useful low-level descriptors including zero-crossing-rate (ZCR), root mean square (RMS) frame energy, pitch frequency and harmonics-to-noise ratio (HNR); standard speech features including mel-frequency cepstral coefficients (MFCCs) and their derivatives; other features including mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE). Up to 1582 acoustic features and transliteration have been investigated.

2. Application of new feature selection method. Correlation based feature subset selection with SFFS was employed to eliminate redundant features because of large feature sets. The experiments proved that spectral features contain the most relevant information about age and gender within speech for almost every pair of age and gender for both databases. When using only LSP features for age and gender recognition, performance was shown to be 6.9% higher compared to the average. Cepstral features performed even 7.1% better than the average feature type. Pitch, as a prosodic Low-Level-Descriptor, prevailed only for the pair male/female where it performed 6.3% better than the average.

3. The use of fuzzy SVM (FSVM) as a new speaker classification method. FSVM assigns a fuzzy membership value as a weight to each training data point. Data points in overlapping regions (consisting of data of different classes) are more important than others. A fuzzy clustering technique is used to determine clusters in these regions. Data points in these clusters will have the highest fuzzy membership value. Fuzzy memberships for other data points are determined by their closest cluster accordingly; therefore their fuzzy membership values will be lower. This means that the decision boundary tends to move to overlapping regions to reduce empirical errors.

4. A detailed comparison of speaker classification performance for GMMs, SVM and FSVM. Different Gaussian components are applied to consider classification

rates for age and gender classification. The one-against-one SVM and FSVM are used for multi-class classification problems.

5. A depth investigation on the relevance of feature type for classification of age and gender. Extensive experiments are performed to determine which features in the speech signal are suited to representation of age and gender in human speech.

6. Classification of age, gender, accent, and emotion characteristics is performed on four well-known data sets including Australian National Database of Spoken Language (ANDOSL), aGender, EMO-DB and FAU AIBO.

1.5 Organisation of the Thesis

This thesis consists of five chapters. Chapter 1 introduces the research project. Chapter 2 reviews current feature extraction, feature selection and classification methods. Fuzzy SVM is introduced in Chapter 3. Chapter 4 presents experimental results and discussions on the use of different features and classification methods. Chapter 5 concludes the thesis and proposes further investigations.

Chapter 2

Literature Review

The aim of this chapter is to provide background knowledge on a speaker classification system and its components. Section 2.1 describes the structure of a speaker classification system. Section 2.2 explains the sound generation process. Section 2.3 explores the extraction of feature vectors from speech signals. Section 2.4 describes the feature selection methods. Finally, Section 2.5 describes the classification techniques used.

2.1 The Speaker Classification System

First we need to differentiate the speaker classification task from the speaker recognition task which includes speaker identification and speaker verification. Speaker identification is the process of determining who is speaking based on information obtained from the speaker's speech. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker classification is the task of assigning a given speech sample to a particular class such as age, gender, accent, or emotion classes. Speaker classification can be thought of as speaker identification in which each class is a speaker. For example, gender classification task can be thought as identifying whether a test utterance is from a male speaker or female speaker.

An automatic speaker classification system includes two phases: training phase and testing phase, see Figure 2.1. In the training phase, the training data of the digital input signal of voice is processed and feature vectors are extracted. Then

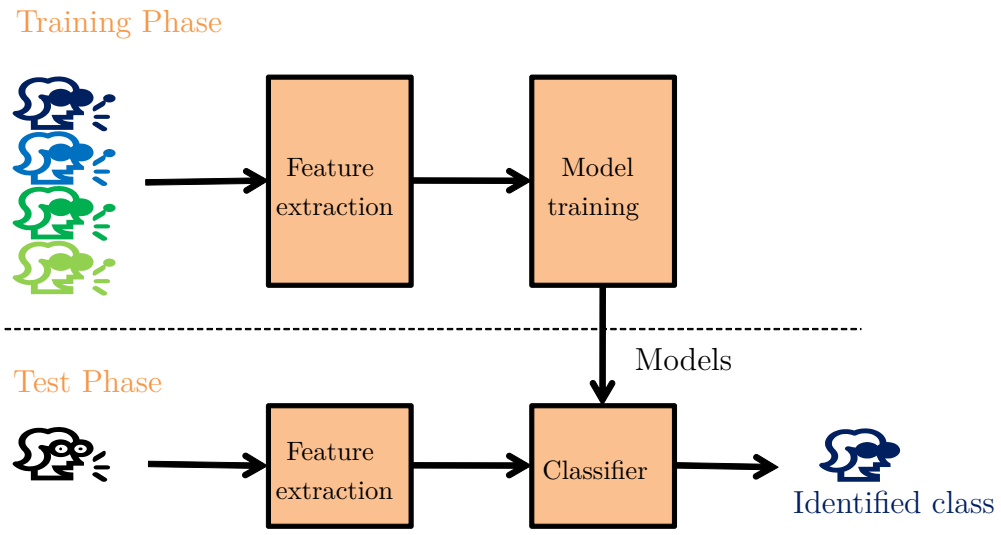

Figure 2.1: Structure of an automatic speaker classification system

these feature vectors of all classes are used to train the speaker class models of a classifier. In the test phase, the input voice signal feature vectors are again extracted. Then they are scored in the classifier to each model and classified into the model given the best score (see Figure 2.2).

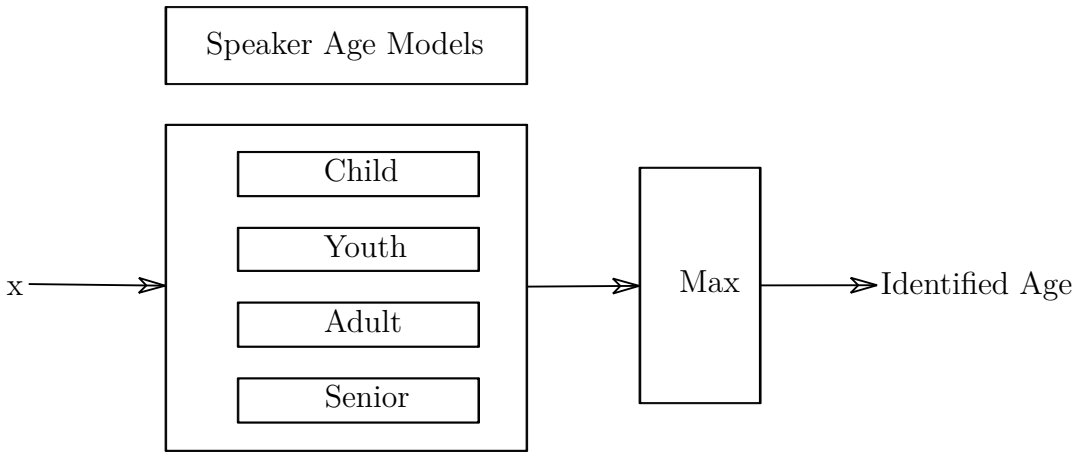

Figure 2.2: Structure of an automatic age classification system

2.2 Sound Generation and Speech Signal

The generation of speech sounds in the vocal tract consists of two processes. In the first process, a constriction in the larynx causes vibration which gives rise to rapid pressure variations. These variations transmit rapidly through the air as sound. In the second process, sound passes through the air cavities of the pharynx, nasal and oral cavities. Sound is changed depending on the shape and size of those cavities. Thus the sound emitted from the lips and nostrils has properties of the sound source and the vocal tract tube. This approach is called the source-filter model of speech production [16].

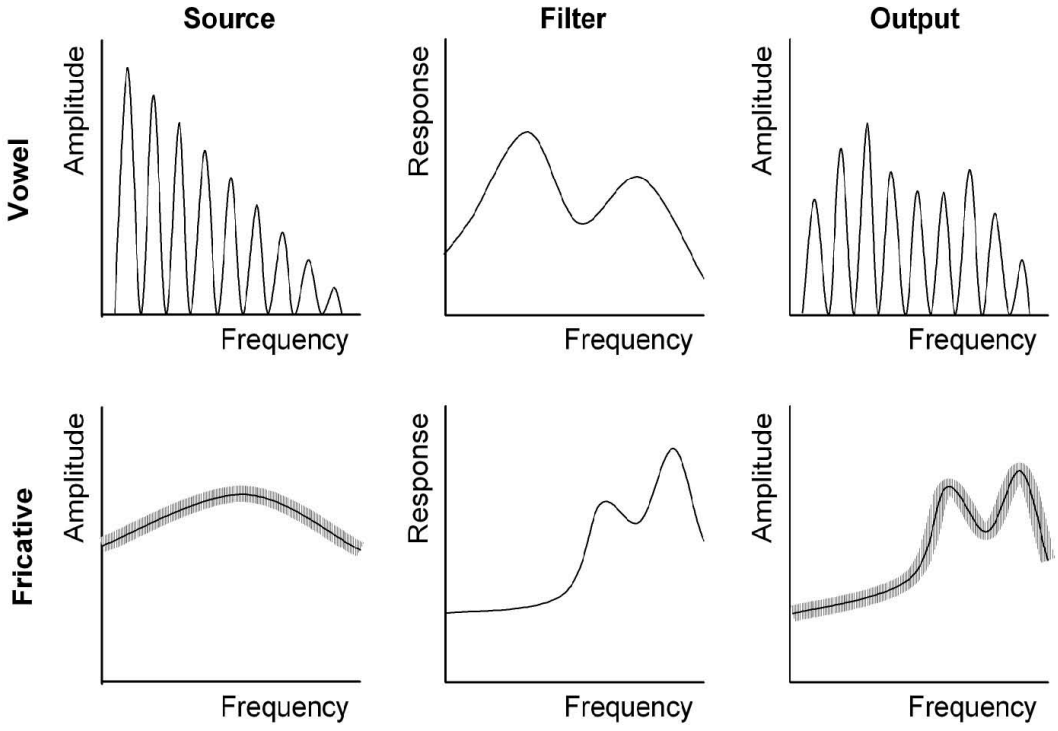

Figure 2.3: Frequency domain diagram of the source-filter explanation of the acoustics of a vowel (voiced) and a fricative (voiceless). The source spectrum (left), the vocal tract transfer function (middle), and the output spectrum (right), after Dellwo [16]

There are two elemental sound generation types: voiced and voiceless, see Figure 2.3. Voiced sounds, also known as phonation, are produced by periodic vibration in the larynx. The vibration happens when sub-glottal pressure increases enough to open

the vocal folds. The air flowing through the glottis causes a decrease in pressure. This closes the folds cutting off the flow and creating a pressure drop above the glottis. The cycle repeats periodically at frequencies between about 50 and 500Hz. The spectrum of this sound is up to about 5000Hz and falling off at about -12dB/octave [16], as shown at top of the left column in the Figure 2.3. Other sound sources are created by turbulence at obstacles to the air-flow. Noise sources caused by the turbulence have broad continuous spectra, varying from about 2 to 6 kHz depending on the exact place and shape of constriction. Normally, noise sources have a single broad frequency peak, rolling off at lower and high frequencies, as shown at the bottom in the left column in Figure 2.3.

The middle column in the Figure 2.3 shows the frequency response of the vocal tract. This frequency response can be modelled by a series of poles called the formants of the tract [16]. The formant frequencies and bandwidths are used as parameters of the vocal tract frequency response. When sound goes out of the lips and nostrils, its frequency shaping is modified again which helps differentiate the signals.

Speech is a time varying signal. In a long period, speech signals are non-stationary but in a short interval between 5 and 100ms, the speech signals are “quasi-stationary”, and the articulatory configuration stays nearly constant. Therefore, speech features are extracted for short frames. The basic mechanisms involved in transforming a speech waveform into a sequence of parameter vectors is illustrated in Figure 2.4. The sampled waveform is analysed in frames with short window sizes so that the signals are “quasi-stationary”. The frames overlap by setting the frame period smaller than the window size. Each frame is then investigated to extract parameters. This process results in a sequence of parameter blocks [70]. SOURCERATE and TARGETRATE in the following figure are the number of samples of the wave source and the number of extracted feature vectors, respectively.

In practice, the window size is typically between 15 ms and 35 ms long with a period of 10 ms. For example, given a waveform sampled at 16kHz and a settings of 30 ms window size with a period of 10 ms, each frame will have 480 samples and will be converted to one feature vector. This results in 100 parameter vectors per second.

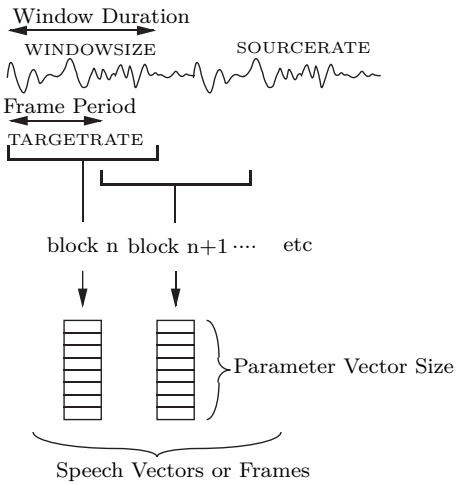

Figure 2.4: Speech encoding process, after Young [70].

We define a frame of speech to be the product of a shifted window with the speech sequence [15]:

$$f_s(n;m) = s(n)w(m-n) \tag{2.1}$$

where $s(n)$ is the speech signal and $w(m-n)$ is a window of length N ending at sample m.

There are some simple pre-processing operations that can be applied before the actual signal analysis. At first, the DC mean (the mean amplitude of the waveform) can be removed from the source waveform [70]. This is useful when the original analogue-digital conversion has added a DC offset to the signal. Second, the signal is usually pre-emphasised by applying the first order difference equation [70]:

$$s'(n) = s(n) - ks(n-1) \tag{2.2}$$

to the samples $s(n), n = 1, \ldots, N$ in each window. Where k in the range $0 \leq k < 1$ is the pre-emphasis coefficient. Finally, the samples in each window usually apply a window with smooth truncations so that discontinuities at the window edges are attenuated [70]. Some of the commonly used windows with smooth truncations are Kaiser, Hamming, Hanning and Blackman. These windows have the benefit of less abrupt

truncations at the boundaries. For Hamming window, the samples $s(n), n = 0, \ldots, N$ in each window apply the following transformation

$$w_n = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right) & 0 \leq n < N \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

2.3 Feature Extraction

This section explores the extraction of feature vectors from speech signals. The large field of speaker classification utilises many properties of spoken language from lower-level features of voice parameters to higher-level features of phonetic, prosodic information. This section presents background knowledge of features generation from low-level to higher-level. Those features are known to carry information about paralinguistic effects includeing energy, pitch (F_0), formants, cepstral, jitter and shimmer, and Harmonics-to-Noise Ratio, resulting in a total of seven feature types that are investigated here. These seven types can be further grouped into three meta-groups: prosodic features, spectral features and voice quality features [55]. The following sections provide a detailed overview.

2.3.1 Spectral features

Spectral features mentioned in this research include Linear Prediction Coding, Formants, Line Spectrum Pair, and Mel-Frequency Ceptral Coefficients.

Linear Prediction Analysis

Linear Prediction Coding is based on a simple model of speech production. The vocal tract is modelled as a set of connected tubes with equal length and piecewise constant diameter. It is assumed that the glottis produces buzzing sounds (voiced speech) or noise (unvoiced speech). Under certain assumptions (no energy loss inside the vocal tract, no nonlinear effects ...) it can be shown that the vocal tract transfer function is modelled by an all-pole filter with the z-transform [70]

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \tag{2.4}$$

where p is the number of poles and $a_0 = 1$. The filter coefficients a_i are chosen to minimise the mean square filter prediction error summed over the analysis window. The autocorrelation method is used to perform this optimisation.

The coefficients of the transfer function are directly related to the resonance frequencies of the vocal tract, called formants, and bear information about the shape of the vocal tract. The coefficients of the transfer function can be directly calculated from the signal through minimizing the linear prediction error [46].

Formants

The formants are related to the vocal tract resonances. The shape and the physical dimensions of the vocal tract decide the location of vocal tract resonances. Speech scientists refer to the resonances as formants because they tend to "form" the overall spectrum. Formant frequencies and bandwidths are important features of the speech spectrum. Formants can be estimated using linear prediction analysis [66].

Line Spectrum Pair

The linear prediction (LP) parameters are rarely used directly. Therefore the line spectrum pair (LSP) was introduced as an alternative in 1980 [15]. These parameters are theoretically equivalent to the LP parameters. But these parameters have smaller sensitivity to quantization noise and have better interpolation properties.

Mel-Frequency Cepstral Coefficients

The filterbank models the ability of the human ear to resolve frequencies non-linearly across the audio spectrum and decreases with higher frequencies. The filterbank is an array of band-pass filters that separates the input signal into multiple components, see Figure 2.5. The filters used are triangular and they are equally spaced along the mel-scale defined by [70]:

$$\text{Mel}(f) = 2595\log_{10}(1 + \frac{f}{700}) \tag{2.5}$$

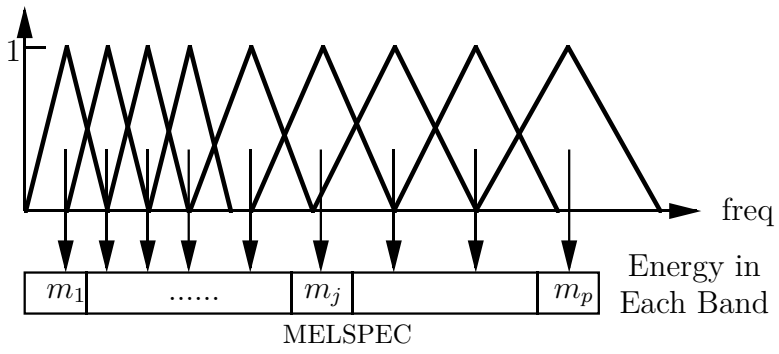

Figure 2.5: Mel-Scale Filter Bank, after Young [70].

Mel-Frequency Cepstral Coefficients (MFCCs) are calculated from the log filterbank amplitudes m_j using the Discrete Cosine Transform

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{n} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{2.6}$$

where N is the number of filterbank channels, c_i are the cepstral coefficients.

2.3.2 Prosodic Features

Timing and rhythms of speech play important roles in the formal linguistic structure of speech communication. Generally prosodic features are related to the tone and rhythm in speech. Since they spread over more than one phoneme segment, prosodic features are suprasegmental. The creation of prosodic features depend on source factors or vocal-tract shaping factors [15]. The source factors are changes in the speech breathing muscles and vocal folds, and the vocal-tract shaping factors relate to the upper articulators movements. Prosodic features include changes in pitch, intensity, and duration.

Pitch

The pitch signal is produced from the vibration of the vocal folds. Two common features related to the pitch signal are the pitch frequency and the glottal air velocity [66]. The vibration rate of the vocal folds is the fundamental frequency of the phonation F_0 or pitch frequency. The air velocity through glottis during the vocal fold vibration is the glottal volume velocity. The most popular algorithm for estimating the pitch signal is based on the autocorrelation [66]. At first, the signal is low filtered at 900 Hz and then it is segmented to short-time frames of speech $f_s(n;m)$. Then the nonlinear clipping procedure that prevents the first formant interfering with the pitch is applied to each frame $f_s(n;m)$ giving

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & \text{if } |f_s(n;m)| > C_{thr} \\ 0 & \text{if } |f_s(n;m)| < C_{thr} \end{cases} \tag{2.7}$$

with C_{thr} is about 30% of the maximum value of $f_s(n;m)$. Next the short-term autocorrelation is determined by

$$r_s(\eta;m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m) \tag{2.8}$$

where η is the lag. Finally, the pitch frequency of the frame ending at m can be given by

$$\hat{F}_0(m) = \frac{F_s}{N} \text{argmax}_\eta \{|r(\eta;m)|\}_{\eta=N(F_l/F_s)}^{\eta=N(F_h/F_s)} \tag{2.9}$$

where F_s is the sampling frequency, and F_l, F_h are the lowest and highest perceived pitch frequencies by humans, respectively. Normally, $F_s = 8000$ Hz, $F_l = 50$ Hz, and $F_h = 500$ Hz [66]. The maximum value of the autocorrelation $\max\{|r(\eta;m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_h/F_s)}$ represents the glottal velocity volume.

Energy

These features model intensity based on the amplitude. The energy is computed as the average of the signal energy, that is, for speech samples $s(n), n = 1, \ldots, N$,

the short-term energy of the speech frame ending at m is [66]

$$E_s(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} f_s(n;m)^2 \tag{2.10}$$

Duration

Duration based features model aspects of temporal lengthening of words [62]. In addition to the absolute duration of a word, two types of normalisation techniques are added to the feature vector. The first is the normalisation of the duration of a word by its number of syllables. The second is the normalisation along the same lines as for the energy normalization. The relative positions on the time axis of energy or pitch features also represent duration because they are measured in milliseconds and were proven to be highly correlated with duration features in [55].

Zero Crossing Measure

The number of zero crossings, or number of times the sequence changes sign, is also a useful feature in speech analysis. The short-term zero crossing measure for the N-length interval ending at $n = m$ is given by [15]:

$$Z_s(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{|sign\{s(n)\} - sign\{s(n-1)\}|}{2} w(m-n) \tag{2.11}$$

where

$$\text{sign}\{s(n)\} = \begin{cases} +1 & \text{if } s(n) \geq 0 \\ -1 & \text{if } s(n) < 0 \end{cases} \tag{2.12}$$

Probability of Voicing

Pitch detection has high accuracy for voiced pitch hypotheses but the performance degrades significantly as the signal condition deteriorates. Pitch extraction for telephone speech is more difficult because the fundamental is often weak or missing. Therefore it is more useful to provide F_0 value and probability of voicing at the same

time. The hypothesis is that first, voicing decision errors will not be manifested as absent pitch values; second, features such as those describing the shape of the pitch contour are more robust to segmental misalignments; and third, a voicing probability is more appropriate than a "hard" decision of 0 and 1, when used in statistical models [10].

2.3.3 Voice Quality Features

Voice Quality features include jitter, shimmer and harmonics-to-noise ratio.

Jitter and Shimmer

Jitter and shimmer are micro fluctuations in vocal fold frequency and amplitude. They are correlated to rough or hoarse voice quality [57]. As shown in Figure 2.6, the major difference is that shimmer has irregular amplitude at regular frequency while in contrast jitter has irregular frequency at regular amplitude. The wave in the top picture has irregular amplitude at the third peak and the wave in the bottom picture has irregular frequency at the second peak.

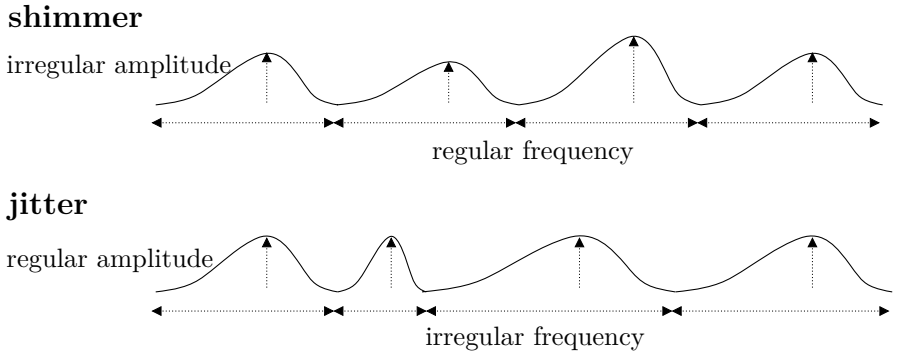

Figure 2.6: Micro variations in vocal fold movements can be measured as shimmer (variation in amplitude) and jitter (variation in frequency), after Schotz [57].

Jitter indicates cycle-to-cycle changes of the fundamental frequency and is approximated as the first derivative of the fundamental frequency [62]. These changes are considered as variations of the voice quality.

$$\text{jitter}(n) = \frac{|F_0(n+1) - F_0(n)|}{F_0(n)} \tag{2.13}$$

where $F_0(n)$ is the fundamental frequency at sample n. Shimmer indicates changes of the energy from one cycle to another.

$$\text{shimmer}(n) = \frac{|\text{en}(n+1) - \text{en}(n)|}{\text{en}(n)} \tag{2.14}$$

where $\text{en}(n)$ is energy of sample n.

Harmonics-to-Noise Ratio

The harmonics-to-noise ratio measures the degree of periodicity of a voiced signal [62]. It can be found from the relative height of the maximum of the autocorrelation function.

2.3.4 Delta and Acceleration Coefficients

The time derivatives to the basic features can help improve the performance of a speaker classification. The delta coefficients are computed using the following regression formula [70]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.15}$$

where d_t is a delta coefficient at time t, c_t is a feature at time t, and Θ is window size. The acceleration coefficients are computed using the same formula onto the delta coefficients.

2.3.5 Static Features

Those features presented above are called low level descriptors (LLD). Static feature vectors are derived per speaker turn by a projection of each uni-variate time series X onto a scalar feature x of real value (R^1) independent of the length of the turn [68].

$$F : X \rightarrow x \in R^1 \tag{2.16}$$

Functional F includes statistical functionals, regression coefficients and transformations are applied to each contour on the turn-level [21, 47].

2.3.6 Discussion

LPC was an efficient method for coding of speech in the 1960s, however MFCCs became the standard feature set in the 1980s and reduced the relevance of LPC features [46]. MFCCs are the choice for many speech recognition applications [70]. They give good discrimination and help a number of manipulations. When applying these frame-level features from the speech recognition to the speaker classification area, it is quite successful in the task of age, gender, dialect, or emotion classification [39, 59, 23]. However those frame based features fails to capture longer-range and linguistic information that also resides in the signal [61].

Higher-level features based on linguistic or long-range information can carry information about paralinguistic effects. Prosodic or suprasegmental features can capture speaker-specific differences in intonation, timing, loudness, pitch [61]. Voice quality features included jitter/shimmer and other measures of micro-prosody, NHR, HNR and autocorrelation reflect the breathiness or harshness in voice [47].

For age classification, acoustic correlates of speaker age are always present in speech. However, the relationships among the correlates are quite complex and are influenced by many factors. For example, there are differences between female and male age, between speakers of good and poor physiological condition, and also between different speech sample types (e.g. sustained vowels, read or spontaneous speech). More research is thus needed in order to build reliable automatic classifiers of speaker age. Some results of acoustic correlates of speaker age have been found [57]. It has been shown that older speakers have a higher variation of acoustic features when compared with young speakers. For example, increased variation has been found in F_0, speech rate, vocal sound pressure level (SPL), jitter, shimmer and HNR. More

differences have been found for male than female speakers, and correlations seem to vary with speech sample type.

For emotion classification, anger is the emotion of the highest energy and pitch level. Ververidis showed the facts in [66] as follows: Angry males show higher levels of energy than angry females. Disgust is expressed with a low mean pitch level, a low intensity level, and a slower speech rate than the neutral state. Fear is correlated with a high pitch level and a raised intensity level. Low levels of the mean intensity and mean pitch are measured when the subjects express sadness. The pitch contour trend is a valuable parameter, because it separates fear from joy. Fear resembles sadness having an almost downwards slope in the pitch contour, whereas joy exhibits a rising slope. The speech rate varies within each emotion. An interesting observation is that males speak faster when they are sad than when they are angry or disgusted. The trends of prosody contours include discriminatory information about emotions.

Table 2.1 gives a summary of the effects of several emotion states on selected acoustic features.

Table 2.1: Summary of the effects of several emotion states on selected acoustic features, after Ververidis [66]. Explanation of symbols: $>$: increases, $<$: decreases, $=$: no change from neutral, $\nearrow$: inclines, $\searrow$:declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: M stands for males and F stands for females.

	Pitch				Intensity		Timing	
	Mean	Range	Variance	Contour	Mean	Range	Speech rate	Transmission duration
Anger	$>>$	$>$	$>>$		$>>_M, >_F$	$>$	$<_M, >_F$	$<$
Disgust	$<$	$>_M, <_F$			$<$		$<<_M, <_F$	
Fear	$>>$	$>$		$\nearrow$	$=>$			$<$
Joy	$>$	$>$	$>$	$\searrow$	$>$	$>$		$<$
Sadness	$<$	$<$	$<$	$\nearrow$	$<$	$<$	$>_M, <_F$	$>$

2.4 Feature Selection

The goal of feature selection (FS) is to select a subset of d features from the given set of D measurements, $d < D$, without significantly degrading (or possibly even improving) the performance of the recognition system [41]. Reducing the dimensionality of the data helps the classification system operate faster and more effectively.

Feature selection algorithms include two broad categories: wrapper methods and filter methods [26]. Wrapper methods use the actual target learning algorithm to estimate the accuracy of feature subsets with a statistical re-sampling technique (such as cross validation). These methods are useful for small data sets but for large data sets they are very slow to execute because the learning algorithm is called repeatedly. On the other hand, filter methods operate independently of any learning algorithm. Redundant features are eliminated before the classification process. Filters usually use all training data when selecting a subset of features.

Correlation-based Feature Selection uses a correlation based heuristic to evaluate features [26]. Although an exhaustive search is necessary to find an optimal subset, in most practical applications this approach is computationally expensive. Therefore research on FS has focused on sequential suboptimal search methods. Among the suboptimal search procedures, the Sequential Floating Forward Selection (SFFS) has proven effective because it can handle high dimensionality involving nonmonotonic criterion by backtracking ability. After each forward step, SFFS applies a number of backward steps as long as the resulting subsets are better than the previous ones [41]. As a result, there are no backward steps if the performance cannot be improved. Thus backtracking in these algorithms is controlled dynamically [41].

SFFS Algorithm

Input:

$Y = \{Y_j | j = 1, \ldots, D\}$ //available measurements//

Output:

$X_k = \{x_j | j = 1, \ldots, k, x_j \in Y\}, k = 0, 1, \ldots, D$

Initialisation:

$X_0 := \emptyset; k := 0$

(in practice one can begin with $k = 2$ by applying SFS twice)

Termination:

Stop when k equals the number of features required

Step 1 (Inclusion)

$x^+ := \arg \max_{x \in Y - X_k} J(X_k + x)$ {the most significant feature with respect to X_k

$X_{k+1} := X_k + x^+; k := k + 1$

Step 2 (Conditional Exclusion)

$x^- := \arg \max_{x \in X_k} J(X_k - x)$ {the least significant feature in X_k

if $J(X_k - \{x^-\}) > J(X_k - 1)$ then

$X_{k-1} := X_k - x^-; k := k - 1$

go to **Step 2**

else

go to **Step 1**

2.5 Classification Methods

This section presents the mathematical modelling techniques for speaker classification including GMMs and SVM.

2.5.1 Gaussian Mixture Models

Speaker classification can be thought as speaker identification in which each class is a speaker. For a reference group of S speaker classes $A = \{1, 2, \ldots, S\}$ represented by models $\lambda_1, \lambda_2, \ldots, \lambda_S$, the objective is to find the speaker class model which has the maximum posterior probability for the input feature vector sequence, $X = \{x_1, \ldots, x_T\}$. The minimum error following Bayes decision rule for this problem

is [43]:

$$\begin{aligned} \hat{s} &= \arg\max_{1\le s\le S} \Pr(\lambda_s|X) \\ &= \arg\max_{1\le s\le S} \frac{p(X|\lambda_s)}{p(X)} \Pr(\lambda_s) \end{aligned} \tag{2.17}$$

Assuming equal prior probabilities of speakers, the terms $Pr(\lambda_s)$ and $p(X)$ are constant for all speakers and can be ignored in the maximum. Using logarithms and the assumed independence between observations, the decision rule becomes

$$\hat{s} = \arg\max_{1\le s\le S} \sum_{t=1}^{T} \log p(x_t|\lambda_s) \tag{2.18}$$

where $p(x_t|\lambda_s)$ is given in Eq. (2.21). The diagram of the speaker classification system is shown in Figure 2.7.

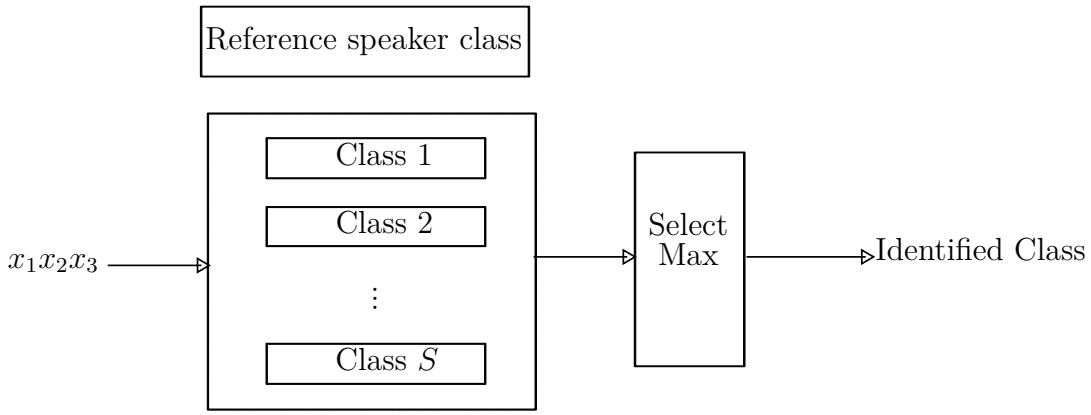

Figure 2.7: Speaker classification system

Since the distribution of feature vectors in X is unknown, it is approximately modelled by a mixture of Gaussian densities, which is a weighted sum of K component densities, given by the equation

$$p(x_t|\lambda) = \sum_{i=1}^{K} w_i N(x_t, \mu_i, \Sigma_i) \tag{2.19}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, $w_i, i = 1, \ldots, K$, are the mixture weights and $N(x_t, \mu_i, \Sigma_i), i = 1, \ldots, K$, are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)'\Sigma_i^{-1}(x_t - \mu_i)\right\}}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \tag{2.20}$$

In training the GMMs, these parameters are estimated such that in some sense, they best match the distribution of the training vectors. The most widely used training method is the maximum likelihood (ML) estimation. For a sequence of training vectors X, the likelihood of the GMMs is

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \tag{2.21}$$

The aim of ML estimation is to find a new parameter model $\bar{\lambda}$ such that $p(X|\bar{\lambda}) \geq p(X|\lambda)$. Since the expression (2.21) is a nonlinear function of parameters in λ its direct maximisation is not possible. However, parameters can be obtained iteratively using the expectation-maximisation (EM) algorithm [29]. An auxiliary function Q is used

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{T} p(i|x_t, \lambda) \log[\bar{w}_i N(x_t, \bar{\mu}_i, \bar{\Sigma}_i)] \tag{2.22}$$

where $p(i|x_t, \lambda)$ is the a posteriori probability for acoustic class $i, i = 1, \ldots, K$ and satisfies

$$p(i|x_t, \lambda) = \frac{w_i N(x_t, \mu_i, \Sigma_i)}{\sum_{k=1}^{c} w_k N(x_t, \mu_k, \Sigma_k)} \tag{2.23}$$

The basis of the EM algorithm is that if $Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda)$ then $p(X|\bar{\lambda}) \geq p(X|\lambda)$ [31, 43]. The following re-estimation equations are found

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} p(i|x_t, \lambda) \tag{2.24}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)} \tag{2.25}$$

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda)(x_t - \bar{\mu}_i)(x_t - \bar{\mu}_i)'}{\sum_{t=1}^{T} p(i|x_t, \lambda)} \tag{2.26}$$

2.5.2 Support Vector Machine

Binary Case

Consider the training data $\{x_i, y_i\}$, $i = 1, ..., n$, $x_i \in R^d$, where label $y_i \in \{-1, 1\}$. The support vector machine (SVM) using C-Support Vector Classification (C-SVC) algorithm will find the optimal hyperplane [8]:

$$f(x) = w^T \Phi(x) + b \tag{2.27}$$

to separate the training data by solving the following optimization problem:

$$\min \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{2.28}$$

subject to

$$y_i \left[w^T \Phi(x_i) + b \right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \ , i = 1, ..., n \tag{2.29}$$

The optimization problem (2.28) will guarantee to maximize the hyperplane margin while minimizing the cost of error, where $\xi_i, i = 1, ..., n$ are non-negative slack variables introduced to relax the constraints of separable data problems to the constraint (2.29) of non-separable data problems as seen in Figure 2.8. For an error to occur the corresponding must exceed unity (see Eq. (2.29)), so $\sum_i \xi_i$ is an upper bound on the number of training errors. Hence an extra cost $C \sum_i \xi_i$ for errors is added to the objective function (see Eq. 2.28) where C is a parameter chosen by the user.

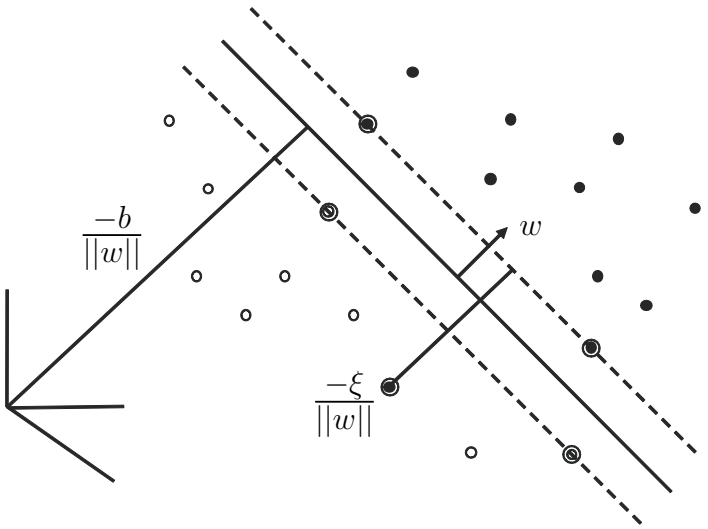

Figure 2.8: Linear separating hyperplane for the non-separable data. The slack variable ξ allows misclassified point.

The Lagrangian formulation of the primal problem is:

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\{y_i({x_i}^T w + b) - 1 + \xi_i\} - \sum_i \mu_i\xi_i \tag{2.30}$$

We will need the Karush-Kuhn-Tucker conditions for the primal problem to attain the dual problem:

$$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \Phi(x_i)^T\Phi(x_i) \tag{2.31}$$

subject to

$$\begin{aligned} 0 \le \alpha_i \le C \\ \sum_i \alpha_i y_i = 0 \end{aligned} \tag{2.32}$$

The solution is given by

$$w = \sum_i^{N_S} \alpha_i y_i x_i \tag{2.33}$$

where N_S is the number of support vectors. Notice that data only appear in the training problem, Eq. (2.30) and Eq. (2.31), in the form of dot product and can be

replaced by any kernel K with $K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j), \Phi$ is a mapping to map the data to some other (possibly infinite dimensional) Euclidean space. One example is Radial Basis Function (RBF) kernel $K(x_i, x_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$

In test phase an SVM is used by computing the sign of

$$f(x) = \sum_{i}^{N_S} \alpha_i y_i \Phi(s_i)^T \Phi(x) + b = \sum_{i}^{N_S} \alpha_i y_i K(s_i, x) + b \tag{2.34}$$

where the s_i are the support vectors.

Multi-class Support Vector Machine

The binary SVM classifiers can be combined to handle the multi-class case: One-against-all classification uses one binary SVM for each class to separate their members from other classes, while one-against-one or pairwise classification uses one binary SVM for each pair of classes to separate members of one class from members of the other. In one-against-one approach, there are $n(n-1)/2$ class pairs decision functions were trained. In test phase, the voting stategy was used as follow: each binary classification was considered to be a voting where votes could be cast for all data points x. The final result was the class with maximum number of votes [12].

2.5.3 Discussion

GMMs have become the dominant approach in both commercial and research systems. It has been used to model distributions of spectral information from short time frames of speech. It can reflect information about a speaker's vocal physiology, and is text-independent because it does not rely on phonetic content [61].

GMMs were effectively used for robust text-independent speaker identification and verification [43, 45]. Gaussian components are capable of modelling underlying acoustic classes representing some broad phonetic events, such as vowels, nasals, or fricatives. These acoustic classes reflect some general speaker-dependent vocal tract configurations. More over, a linear combination of Gaussian densities is capable of representing a large class of sample distributions. The mean component density can

represent the spectral shape of an acoustic class, and the covariance matrix can represent variations of the average spectral shape. An important problem of GMMs are how to determine the number of components in a mixture needed because there is no theoretical way to find out it. This number should be chosen adequately to model a speaker class and be as small as possible to guarantee performance [45].

Chapter 3

Proposed Methods

The purpose of this study has three main parts. The first part is to derive fuzzy SVM (FSVM) developed as an extension of SVM. The second part is to compare the performance of GMMs and that of SVM. The third part is to improve the accuracy of speaker classification by applying FSVM and to investigate the relevance of feature type for classification of age and gender. These studies are conducted on four well-known datasets of age, gender, accent, and emotion characteristics. The structure of this chapter is as follows. Section 3.1 presents the FSVM method. Section 3.2 presents accent classification based on frame-level features using GMMs and static features using SVM. Section 3.3 investigates classification of speaker characteristics based on higher-level features using GMMs, SVM and FSVM. Section 3.4 explores the relevance of feature type for classification of age and gender.

3.1 Fuzzy Support Vector Machine

Fuzzy SVM is modelled as follows

$$\min_{w,b} \quad (\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} {\lambda_i}^{\beta}\xi_i) \tag{3.1}$$

subject to

$$\begin{aligned} & y_i \left[\mathbf{w}^T \phi(\mathbf{x}_i) + b\right] \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{3.2}$$

where weights $\lambda_i \in [0, 1], i = 1, \ldots, n$ are regarded as fuzzy memberships and $\beta > 0$ is a parameter to slightly adjust the membership function in overlapping region. This approach assumes that training data points should not be treated equally to avoid the problem of sensitivity to noise and outliers. The corresponding dual form is as follows

$$\min_{\alpha} (\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i) \tag{3.3}$$

subject to

$$\begin{aligned} & 0 \leq \alpha_i \leq {\lambda_i}^{\beta} C \quad i = 1, ..., n \\ & \sum_{i=1}^{n} y_i \alpha_i = 0 \quad i = 1, ..., n \end{aligned} \tag{3.4}$$

The same decision function is used: $f(x) = \text{sign}(w^T \phi(x) + b)$.The unknown data point x belongs to positive class if $f(x) = +1$ or negative class if $f(x) = -1$.

3.1.1 Calculating Fuzzy Memberships

A simple yet efficient method is proposed to determine fuzzy memberships. The positive and negative data points are normally overlapped and the task of fuzzy SVM is to construct a hyperplane in feature space to separate positive data from negative data. Hence we assume that the data points in the overlapping regions are important and they should have the highest fuzzy membership value. Other data points are less important and should have lower fuzzy membership values.

3.1.2 Fuzzy Clustering Membership

Fuzzy clustering membership is determined using the algorithm below. In step 1, a clustering algorithm is chosen, for example fuzzy c-means clustering in this research. In step 2, the chosen clustering algorithm is run on training data set to determine

separated data clusters. In step 3, clusters that contain both positive and negative data are determined and considered as the overlapping regions. In step 4, fuzzy memberships of data points in these overlapping regions are set to 1, highest membership. In step 5, fuzzy memberships of other data points are determined by their closest cluster accordingly. Although clustering is performed in the input space, according to most current kernel functions, relative distances between data points are preserved so we can apply the clustering results obtained in the input space to the feature space.

Fuzzy Membership Calculation Algorithm

Step 1. Select a clustering algorithm

Step 2. Perform clustering on the training data set

Step 3. Determine a subset containing clusters that contain both positive and negative data. Denote this subset as $MIXEDCLUS$.

Step 4. For each data point $x \in MIXEDCLUS$, set its fuzzy membership to 1

Step 5. For each data point $x \notin MIXEDCLUS$, do the following

a. Find nearest cluster to x

b. Calculate fuzzy membership of x to this cluster

3.1.3 The Role of Fuzzy Memberships

The term $\sum_i {\lambda_i}^{\beta}\xi_i$ is regarded as a weighted sum of empirical errors to be minimized in fuzzy SVMs. If a misclassified point x_i is not in a mixed cluster, its fuzzy membership λ_i is small and hence its error ξ_i can be large, as long as ${\lambda_i}^{\beta}\xi_i$ is still minimized, as in Figure 3.1. On the other hand, if it is in a mixed cluster, its fuzzy membership is 1 and hence its error ξ_i must be small such that ${\lambda_i}^{\beta}\xi_i$ remains minimized. This means that the decision boundary tends to move to overlapping regions to reduce empirical errors in this region.

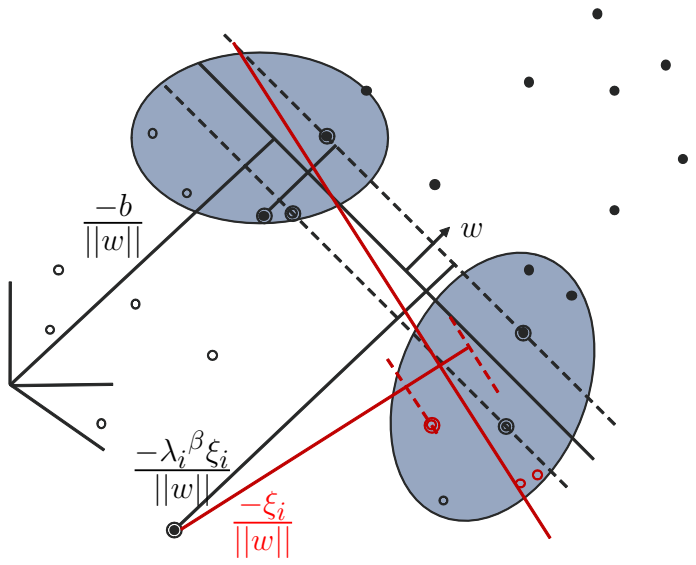

Figure 3.1: Linear separating hyperplanes of SVM and FSVM for the non-separable data. The small membership λ_i allows large error of misclassified point outside overlapping regions, hence the decision boundary tends to move to overlapping regions to reduce empirical errors in this region.

3.2 Speaker Classification using Frame-level Features

For frame-level features, MFCCs are the most commonly used features in modern speaker recognition systems [44]. MFCCs have become the standard feature set for various speech applications. Although originally developed for speech recognition, many state-of-the-art systems for speaker classification use MFCCs as features [24].

Meanwhile, the GMMs approach is a well-known modelling technique in text-independent speaker recognition systems for frame-based features [63]. The Gaussian components are capable of representing characteristic spectral shapes (vocal tract configurations) which comprise a person's voice. That means GMMs can model the underlying acoustic classes of the speakers and the short-term variations of a person's voice. Therefore GMMs can achieve high identification performance for short utterances. GMMs is also considered as a nonparametric, multivariate probability density function model, and it can represent arbitrary feature distributions [43, 45].

Experiments using GMMs and frame-level features on EMODB and ENTERFACE data sets were carried out by Vlasenko and Schuller [67, 53]. Speech signals were processed to obtain 12 MFCCs, log frame energy plus speed and acceleration coefficients to form 39 dimensional feature vectors. Additionally, Cepstral Mean Subtraction (CMS) and variance normalization were also applied. An experiment using 512-mixture, full-covariance GMMs and frame-level features on aGender data set was carried out by Gajsek [23]. In this data set, 12 MFCCs and short-time energy plus speed were extracted from the waveforms. In addition, Cepstral Mean Subtraction (CMS) and variance normalization are also applied. Silent regions were detected and removed by inspecting short-time energy. Experiments using GMMs and frame-level features on AIBO data set were carried out by Schuller [51] as baseline results for the INTERSPEECH 2009 emotion challenge. In detail, the 16 low-level descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by autocorrelation function, and MFCCs 1-12 in full accordance with HTK-based computation.

In this research study, experiments using GMMs and frame-level features were carried out on ANDOSL for accent classification. As stated in the introductory chapter, the Australian accent has a great deal of cultural credibility. It is disproportionately used in advertisements and by newsreaders. Current research on Australian accent and dialect is focusing on a linguistic approach to dialect of phonetic study [28, 5], classification of native and non-native Australian [34], or to improve Australian automatic speech recognition performance [7, 2]. However, there is no research on automatic speaker classification based on the three Australian accents of Broad, General, and Cultivated.

Accent is particularly known to have a detrimental effect on speech recognition performance. By applying higher-level information derived from phonetics rather than solely from acoustics, speaker idiosyncrasies and accent-specific pronunciations can be better covered. Since this information is provided from complementary phone recognizers [56], I anticipate greater robustness, which is confirmed by my results .

3.3 Speaker Classification using Static Features

GMMs with frame-level features are found to be challenged by mismatching acoustic conditions. To overcome these problems, higher-level features based on linguistic or long-range information have been recently investigated [47, 61] . Prosodic and voice quality features are highly correlated to emotion [14, 55]. System in state-of-the-art illustrates that higher-level system outperforms standard systems and provide increasing relative gains as training data increases [61]. These features together are called low level descriptor (LLD) [50]. The higher success of static feature vectors derived by projection of the low level descriptor (LLD) by descriptive statistical functional application such as lower order moments (mean, standard deviation) or extrema is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech [51].

Experiments were carried out on four data sets in this research study. In the first step, feature vectors were extracted from speech signal. For age and gender classification on aGender and ANDOSL data sets, the INTERSPEECH 2010 Paralinguistic challenge 450-feature set was used. For emotion classification on FAU AIBO and EMO-DB, the INTERSPEECH 2009 Emotion challenge 384-feature set was used. Features are extracted using the open-source Emotion and Affect Recognition toolkit's feature extracting backend openSMILE [21]. In the second step, another version of each of these four data sets with an additional feature selection step applying onto these feature sets was created, resulting in a reduced feature set for each data set. The feature selection algorithm chosen was sequential forward floating search (SFFS). In the third step, both the full and reduced feature vectors were converted into HTK format for running GMMs using HTK toolkit, and converted to LIBSVM format for running SVM and FSVM using LIBSVM tool with my extension. For the final step, experiments using GMMs, SVM and FSVM were carried out on those four data sets with and without feature selection. I used SVM and FSVM with one-against-one for multi-class classification problems, i.e. $n(n-1)/2$ class pairs decision functions were trained and a test vector was classified into a class by voting strategy.

All test-runs were carried out in 5-fold cross validation manner for ANDOSL, FAU AIBO, and EMO-DB database, except for the aGender database since it had separated training and developing sets already. At first, the database was separated to 5 folds. Next, a fold was considered as the validation set and the rest were training set.

3.4 Feature Type Relevance in Age and Gender Classification

Features related to speech rate, sound pressure level (SPL) and fundamental frequency (F_0) have been studied extensively, and appear to be important correlates of speaker age. The relationships among the correlates appear to be rather complex, and are influenced by several factors. For instance, differences have been reported between correlates of female and male ages, between speakers of good and poor physiological conditions, between chronological age and perceived age, and also between different speech sample types [57]. Speaker age is a characteristic which is always present in speech. Previous studies have found numerous acoustic features which correlate with speaker age. However, few attempts have been made to establish their relative importance. Many acoustic features of speech undergo significant change with ageing. Earlier studies have found age-related variation in duration, fundamental frequency, SPL, voice quality and spectral energy distribution (both phonatory and resonance). Moreover, a general increase of variability and instability, or instance in F_0 and amplitude, has been observed with increasing age [58].

This resaerch study groups features into six groups:

1. MFCCs [0-14]
2. Log Mel Frequency Band [0-7]
3. LSP Frequency [0-7]
4. PCM loudness

5. Pitch related (F0, F0 Envelope, and Voicing Probability)

6. Jitter and Shimmer (Jitter local, Jitter consecutive frame pairs, Shimmer local)

For each of these groups, classification results using SVM and FSVM are reported for the full feature sets and for the reduced feature sets.

Opposing related speech recognition tasks, the predominant question of optimal features is still an open issue for recognition of affect [55]. Prosodic and voice quality features have been shown useful to speaker characteristics [55, 52]. However, it is not fully investigated and determined which features contribute most in speaker classification. This research attempts to answer this question. Effects of features on speaker classification were investigated on the above-mentioned four data sets.

Chapter 4

Experimental Results

This chapter presents experimental results for speaker classification. Section 4.1 describes data sets used in the experiments. Section 4.2 presents accent classification results using GMMs with MFCCs features on ANDOSL. Section 4.3 presents classification results of age, gender, and emotion on ANDOSL, aGender, EMO-DB, and FAU AIBO data sets. The age and gender feature set and emotion feature set with and without feature selection are employed. Section 4.4 presents the relevance of feature type for the classification of age and gender on ANDOSL and aGender data sets.

4.1 Data Sets

This section describes briefly the four data sets used in the experiments. Since not many age, gender, accent, and emotion data sets were made public, I carried out research on data sets that were available including ANDOSL, aGender, EMODB, eNTERFACE and AIBO. Therefore the number of speaker characteristics is limited in these data sets including age, gender, accent and emotion. However, these data sets are popular and large enough to conduct research on popular speaker characteristics and compare to published results of other researchers. The presented methods can be used for other data sets.

4.1.1 ANDOSL

The Australian National Data set of Spoken Language (ANDOSL) corpus [38] comprised carefully balanced material for Australian speakers, both Australian-born and overseas-born migrants. The aim was to represent as many significant speaker groups within the Australian population as possible. Current holdings were divided into those from native speakers of Australian English (born and fully educated in Australia) and those from non-native speakers of Australian English (first generation migrants having a non-English native language). A subset used for speaker classification experiments in this research study consisted of 108 native speakers. There were 36 speakers of General Australian English, 36 speakers of Broad Australian English and 36 speakers of Cultivated Australian English in this subset. Each of the three groups comprised six speakers of each gender in each of three age ranges (18-30, 31-45 and 46+). So there were 18 groups of 6 speakers labeled as ijk, where i denotes f (female) or m (male), j denotes y (young) or m (medium) or e (elder), and k denotes g (general) or b (broad) or c (cultivated). For example, the group fyg contains 6 female young general Australian English speakers. Each speaker contributed in a single session, 200 phonetically rich sentences. All waveforms were sampled at 20 kHz and 16 bits per sample.

4.1.2 aGender

The aGender corpus [52] was collected by the German Telekom. The subjects repeated given utterances or produced free content prompted by an automated Interactive Voice Response System. The recordings repeated six sessions with one day break in each session to ensure more variations of the voices. The subjects used mobile phone and alternate indoor and outdoor to obtain different recording environments. The associated age cluster was compared with a manual transcription of the self stated date of birth to validate the data. The caller was connected by mobile network or ISDN and PBX to the recording system, which consisted of an application server hosting the recording application and a VoiceXML telephony server (Genesys Voice

Platform). The utterances were stored on the application server as 8 bit, 8 kHz, A-law. All age groups have equal gender distribution. Each of the six recording sessions contained 18 utterances. In total, 47 hours of speech in 5364 single utterances of 954 speakers were collected. The mean utterance length was 2.58 sec. The corpus was randomly divided into three sets of the seven classes, 40%/30%/30% Train/Develop/Test distribution. The Test set included 25 speakers per class (17 332 utterances, 12.45 hours), the Train set (32527 utterances in 23.43 hours of speech of 471 speakers), and the Develop set (20549 utterances in 14.73 hours of speech of 299 speakers). These 7 groups were combined into age group C, Y, A, S or gender group f, m, x, where f and m stand for female and male, and x represents children without gender discrimination as gender discrimination of children is considerably difficult (see Table 4.1).

Table 4.1: Age and gender classes of the aGender corpus, where f and m abbreviate female and male, and x represents children without gender discrimination. The last two columns represent the number of speakers/instances per set.

Class	Group	Age	Gender	# Train	#Develop
1	CHILD	07-14	x	68/4406	38/2396
2	YOUTH	15-24	f	63/4638	36/2722
3	YOUTH	15-24	m	55/419	33/2170
4	ADULT	25-54	f	69/4573	44/3361
5	ADULT	25-54	m	66/4417	41/2512
6	SENIOR	55-80	f	72/4924	51/3561
7	SENIOR	55-80	m	78/5549	56/3826

4.1.3 EMO-DB

The EMO-DB corpus or Berlin Emotional Speech Data set [9] contains recordings of ten professional actors (5 female and 5 male). Each actor simulated the 7 emotions (neutral, anger, fear, joy, sadness, disgust, and boredom) with text that could be

used in everyday communication and are interpretable in all applied emotions. For each emotion, 10 German utterances (5 short and 5 longer sentences) were recorded in an anechoic chamber with high-quality recording equipment. In total, there were 800 utterances (7 emotions * 10 actors * 10 sentences + some second versions). In a perception test judged by 20 listeners, utterances recognised better than 80% and judged as natural by more than 60% of the listeners were phonetically labelled in a narrow transcription with special markers for voice-quality, phonatory and articulatory settings and articulatory features. The data set was recorded in 16 bit, 16 kHz under studio noise conditions. For experiments in this thesis, only the data sets with 60% of the annotators agreeing upon naturalness and 80% upon assignability to an emotion were chosen in accordance to [54]. This final class distribution is shown in Table 4.2.

Table 4.2: Distribution of emotions, data set EMO-DB

	anger (W)	boredom (L)	disgust (E)	fear (A)	happiness (F)	neutral (N)	sadness (T)	Σ
#	127	79	38	55	58	78	53	488

4.1.4 AIBO

The AIBO corpus [51] includes recordings of German children interacting with Sony's pet robot Aibo. The children were led to believe that the Aibo was responding to their commands whereas the robot was actually controlled. Sometimes the Aibo disobeyed commands, thereby provoking emotional reactions. The data was collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz). The recordings were segmented automatically into 'turns' using a pause threshold of 1s. Five labellers (advanced students of linguistics) listened to

the turns in sequential order and annotated each word independently of each other as neutral (default) or as belonging to one of ten other classes. The data was labelled on the word level with majority voting. There were 10 classes containing 48,401 words, in which 4,707 words had no majority voting. For the emotion challenge [51], the 18,216 manually defined chunks based on syntactic-prosodic criteria were used because of the best performance on chunk unit. There were two classification problems. The five-class classification problem covered classes Anger (subsuming angry, touchy, and reprimanding) Emphatic, Neutral, Positive (subsuming motherese and joyful), and Rest were to be discriminated. The two-class problem covered classes NEGative (subsuming angry, touchy, reprimanding, and emphatic) and IDLe (consisting of all nonnegative states). The classes were highly unbalanced (see Table 4.3). The training data was taken from one school (OHM, 13 male, 13 female) and the testing data was taken from the other school (MONT, 8 male, 17 female) to guarantee speaker independence.

Table 4.3: Number of instances for the 5-class problem

#	A	E	N	P	R	Σ
train	881	2093	5590	674	721	9959
test	611	1508	5377	215	546	8257
Σ	1492	3601	10967	889	1267	18216

4.2 Accent Classification

The accent classification experiment was carried out on ANDOSL using GMMs with MFCCs features and SVM with static features.

4.2.1 Parameter Settings for GMMs

GMMs were trained and tested using hidden Markov model toolkit (HTK) which is used for building hidden Markov models (HMMs) [69]. The reason for using HTK

is that GMMs can be seen as one-state continuous HMM. MFCCs features were extracted from speech signals using HTK. The speech data were processed in 32 ms frames at a frame rate of 10 ms. Periods of silence are removed prior to feature extraction by using an automatic energy-based speech/silence detector [70]. Frames were Hamming windowed and pre-emphasised with $m_p = 0.97$. The basic feature set consisted of 12th-order MFCCs and the normalised short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames.

GMMs were initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e. $[\sigma_k]_{ii} = \sigma_k^2$ and $[\sigma_k]_{ij} = 0$ if $i \neq j$, where $\sigma_k^2, 1 < k < K$ are variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [45]. This constraint places a minimum variance value $\sigma_{\min}^2 = 10^{-2}$ on elements of all variance vectors in the GMMs in our experiments.

Performance of GMMs with respect to the number of Gaussian components was explored. The number of components chosen in a GMMs is 16, 32, 64, 128, or 256. The objective is to choose a minimum adequate number of components necessary for a good model while guaranteeing affordable computational complexity both in training and classification [45].

Figure 4.1 presents the classification rate averaged on 10 experiments where the 20 training utterances were randomly selected. Overall the classification rates are higher when the number of Gaussian components increases. The Cultivated accent gets better results for 15 Gaussians or higher and achieves the highest classification rate of 96% for 256 Gaussians.

The standard deviation (STDEV) was measured to consider how widely values are dispersed from the average value. Low STDEV indicates that the values tend to be very close to the mean and the accuracies are consistent when repeating experiments. Table 4.4 shows the STDEV of the accent classification rates for the 10 experiments. The results are consistent for 256 Gaussians.

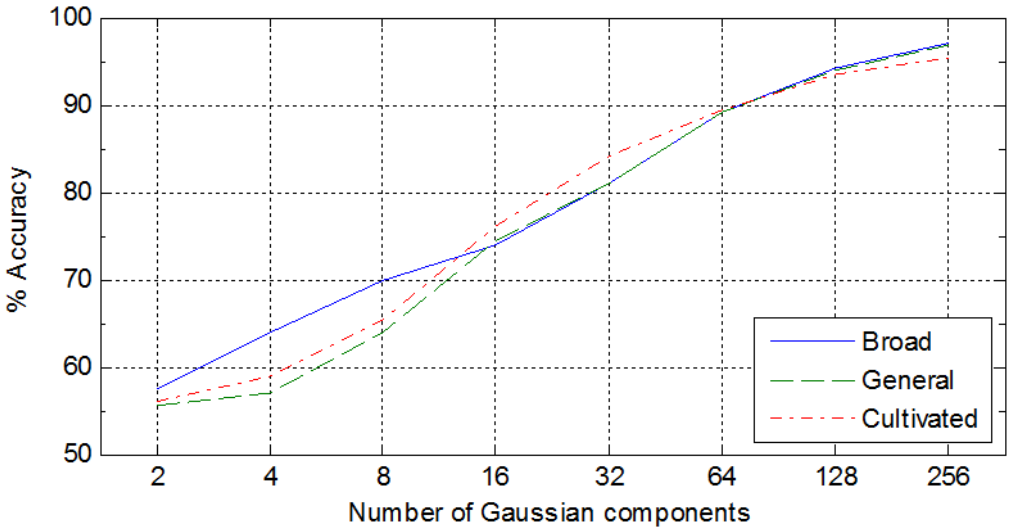

Figure 4.1: Accent classification for Broad, General and Cultivated groups.

Table 4.4: Standard deviation (%) of ACCENT classification from 10 experiments

	Number of Gaussian components							
	2	4	8	16	32	64	128	256
Broad	1.61	2.39	2	1.55	1.24	0.82	0.52	0.34
General	2.91	3.65	4.27	2.74	2.13	1.34	0.92	0.62
Cultivated	2.62	2.09	3.76	2.21	1.51	1.02	0.65	0.61

4.2.2 Parameter Settings for SVM

Experiments were performed using WEKA data mining tool [27], SVM with RBF kernel were selected. All feature vectors were scale to range [-1, 1] in order to avoid domination of some dimension to general performance of classifiers. We performed several experiments with different values of parameters C and γ to search for the best model. The chosen values were $C = 2^1, 2^3, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^3$. The 10-fold cross-validation was used with every pair of values of C and γ. Results are shown in Figure 4.2 and we can see that the best values are $C = 2^7$ and $\gamma = 2^{-5}$ with accuracy of 98.7%.

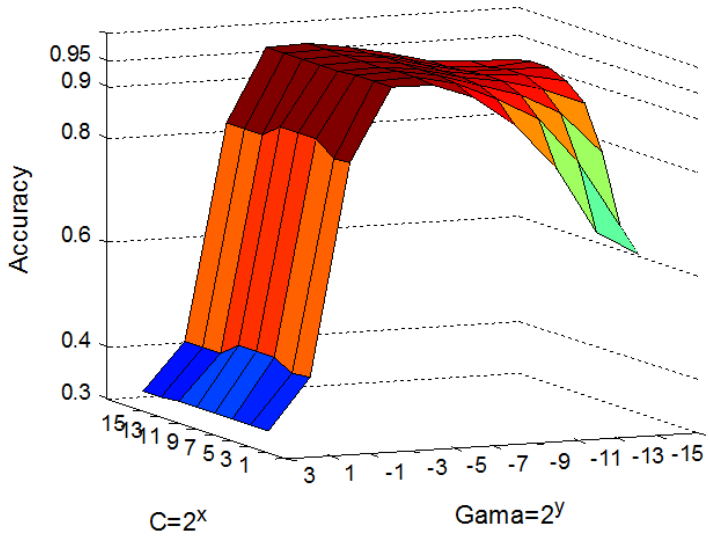

Figure 4.2: Accent classification rates versus C and γ.

4.2.3 Accent Classification Results Versus Age

The influence of age and gender on the accent classification was considered by dividing 108 speakers in to 18 speaker groups based on the three accents Broad, General, and Cultivated, three ages Young, Middle, and Elderly, and two genders Male and Female. Each group contained 6 speakers. The number of Gaussians was set to 256. Figure 4.3 shows the accent classification versus age. While the classification

rates of Broad and General slightly increase from 94% and 96% at Young age to 98% and 97%, respectively for Middle age and Elderly age, the accuracy of Cultivated drops down from 98% for Young to 94% for Middle and to 89% for Elderly. The best results were found for the Middle group. These show that the accent is best recognized for middle speakers and the Cultivated accent is hard to recognize for elderly speakers using GMMs with MFCCs features but it is better using SVM with static features.

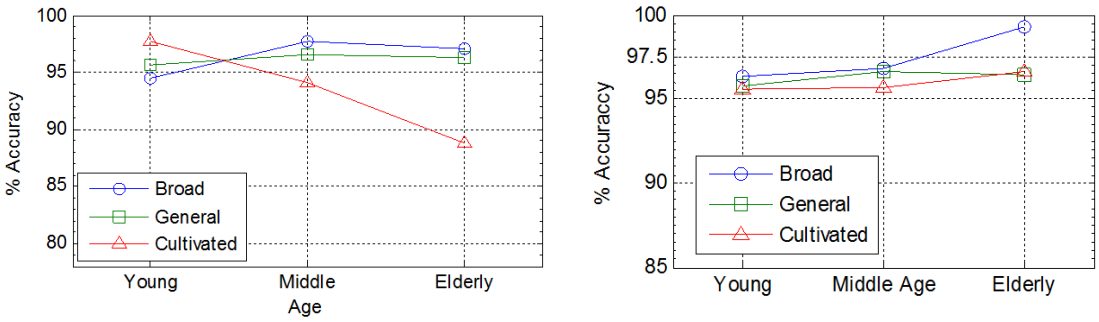

Figure 4.3: Accent classification versus age

Results showed that SVM achieved the highest performance for Accent classifications. As seen in the introductory chapter, Cultivated Australian English has some similarities to British Received Pronunciation, and is often mistaken for it. Both results in Figure 4.3 also show that Cultivated classification achieved the lowest classification rate comparing with the other two accents Broad and General.

4.2.4 Accent Classification Versus Age and Gender

Figure 4.4 and Figure 4.5 show the accent classification rates versus age performed on male and female speakers, respectively. The Cultivated accent is recognizable on male speakers only.

Similar to the previous classification results of GMMs, we also considered the robustness of this classification by calculating the STDEV values on 10 experiments and listed them on Table 4.5 below. The values are very low, ranging from 0.39% to 1.41%, which guarantee the robustness.

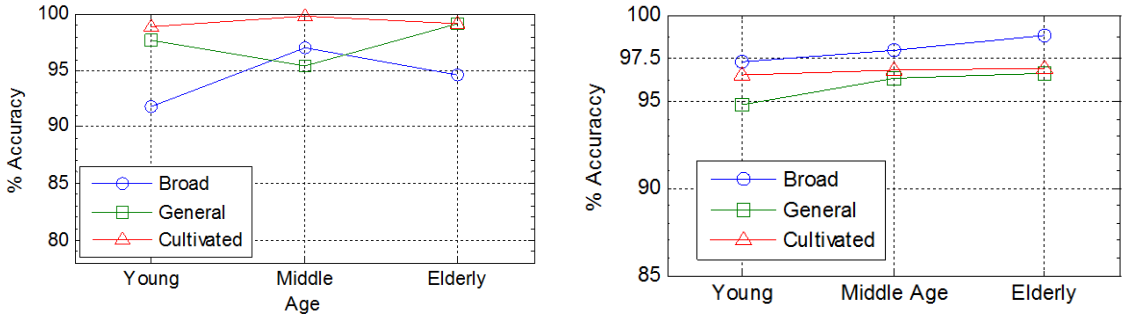

Figure 4.4: Accent classification versus age performed on male speakers

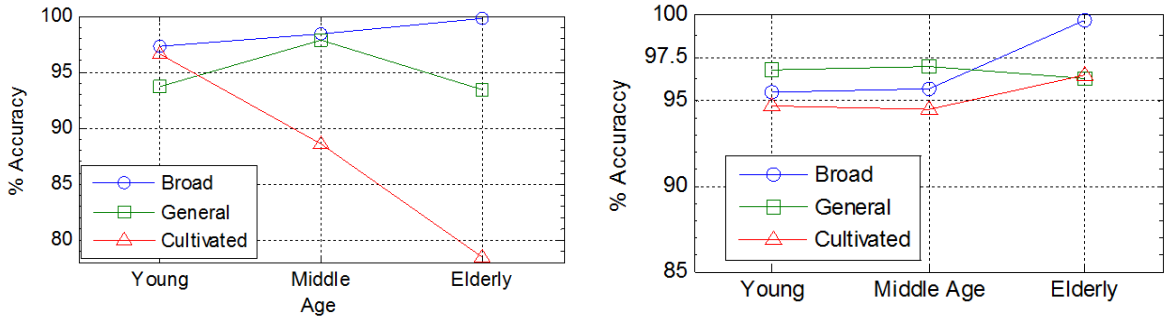

Figure 4.5: Accent classification versus age performed on female speakers

Table 4.5: Standard Deviation (%) of Accent classification Accuracy Versus Age Averaged on 10 experiments

	Age group			Age group of Male			Age group of Female		
	Y	M	E	Y	M	E	Y	M	E
Broad	0.79	0.6	0.51	1.52	0.85	0.98	0.5	0.67	0.16
Cultivated	0.48	1.41	0.39	0.68	0.14	0.4	0.81	2.73	0.78
General	1.16	0.36	1.02	0.99	0.68	0.35	2.26	0.48	1.91

4.3 Age, Gender and Emotion Classification Using Static Features

Experiments of Age, Gender, and Emotion classification using static features were performed. For the Age and Gender classification task on aGender and ANDOSL data sets, the INTERSPEECH 2010 Paralinguistic challenge 450-feature set was used [52]. For the Emotion classification task on FAU AIBO and EMO-DB, the INTERSPEECH 2009 Emotion challenge 384-feature set was used [51].

The Paralinguistic features included 450 features of the INTERSPEECH 2010 paralinguistic challenge [52]. These features include (see Table 4.6) 29 low-level descriptors, their first order regression coefficients, 8 functionals and the 2 single features F_0 number of onsets and turn duration. There are 16 features eliminated due to zero-information.

Table 4.6: Paralinguistic feature set for Age and Gender classification, after Schuller [52].

Descriptors	**Functionals**
MFCCs 0-14	arithmetic mean
LSP Frequency 0-7	standard deviation
F_0	skewness
F_0 Envelope	kurtosis
Voicing Probability	percentile 1/99
Jitter local	percentile range 99-1
Jitter consecutive frame pairs	
Shimmer local	

The Emotion features consisted of 384 features of the INTERSPEECH 2009 emotion challenge [51]. This feature set included the most common and promising feature types and functionals for emotion classification [51]. The 384 features were extracted

in three step as before: at first, 16 low-level descriptors (see Table 4.7) were extracted. Next, their first order regression coefficients were added. Then 12 functionals (see Table 4.7) were applied.

Table 4.7: Emotion feature set for Emotion classification, after Schuller [51].

Descriptors	**Functionals**
Zero-crossing-rate	mean
Root mean square Energy	standard deviation
F_0	kurtosis, skewness
Harmonics-to-noise ratio	extremes: value, rel. position, range
MFCCs 1-12	linear regression: offset, slope, MSE

At the first stage, the above-mentioned feature vectors were extracted from speech signal of the four data sets aGender, ANDOSL, FAU AIBO, and EMO-DB using the open-source Emotion and Affect Recognition toolkit's feature extracting backend openSMILE [21]. At the second stage, feature selection algorithm was run on another version of each of those four data sets, resulting in a reduced feature set for each data set. The chosen feature selection algorithm was sequential forward floating search (SFFS). The remaining features after SFFS was shown in the appendices section. At the final stage, experiments using SVM and FSVM were carried out on those four data sets, with and without feature selection. Since SVM technique constructed only one discriminant function to separate two classes, in order to handle multi-class classification problems, the one-against-one approach was used [12], i.e. $n(n-1)/2$ class pairs decision functions were trained. In the test phase, the voting stategy was used as follows: each binary classification was considered to be a voting where votes could be cast for all data points x. The final result was the class with maximum number of votes.

All test-runs were carried out in 5-fold cross validation manner for ANDOSL, FAU AIBO, and EMO-DB data set, except for the aGender data set since it had separated

training and developing sets already. At first, the data set was separated into 5 folds. Next, a fold was considered as the validation set and the rest were the training set. The cross validation accuracy was the average of accuracies on predicting the validation sets [12]. Parameter selection procedures were employed to find the best parameters as follows: a possible interval of C and γ for SVM or additional β for FSVM were chosen as the grid space, $C = 2^1, 2^3, \ldots, 2^{15}$, $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^3$, $\beta = 2^{-2}, 2^{-1}, \ldots, 2^3$. Then, all grid points of (C, γ) for SVM or (C, γ, β) for FSVM were tried to find which one gave the highest cross validation accuracy. The best parameter set was then selected.

For evaluation measures I used unweighted accuracy (UA) as the main measure to better reflecting unbalance among classes and weighted accuracy (WA) as an extra measure. Where the number of instances per class was not distributed equally, the results for classes with fewer instances tended to be below the actual achievable result for that class. This is due to insufficient amount of training data and low a-priori probability for the class [50]. Similarly, the classes with more instances have a higher a-priori probability and a more robustly estimated model due to the higher amount of available training data. The recognition performance for such a class is usually over-estimated. Classes with many instances have a substantial influence on the WA measure, whereas smaller classes have very little influence. Thus, the WA measure tends to over-estimate performance when the instance numbers of the classes are very unbalanced. Likewise, not weighting the mean recall rate with the number of instances in each class increases the influence of classes with very few instances, thus under-estimating overall performance [55].

Table 4.8 shows the classification rates of SVM and FSVM on all four data sets with full feature sets. Overall, FSVM shows slighly better performance than SVM, 0.05% in FAU AIBO data set, or even shows the same performance as SVM in ANDOSL data set.

Table 4.9 shows the classification rates of SVM and FSVM on all four data sets with SFFS feature selection. Overall, the accuracies were a decrease as a result of feature reduction. Using SVM, ANDOSL was nearly unaffected, aGender was reduced

Table 4.8: Classification rates (%) of SVM and FSVM on the four data sets.

	SVM			FSVM		
	UA	WA	Parameters (C, γ)	UA	WA	Parameters (C, γ, β)
ANDOSL	99.2	99.2	$(2^5, 2^{-5})$	99.2	99.2	$(2^5, 2^{-5}, 2^{-1})$
aGender	45.09	45.46	$(2^7, 2^{-9})$	45.11	45.47	$(2^5, 2^{-9}, 2^1)$
FAU AIBO	45.89	63.59	$(2^5, 2^{-5})$	45.94	63.57	$(2^5, 2^{-5}, 2^2)$
EMO-DB	81.89	83.81	$(2^3, 2^{-7})$	81.91	83.6	$(2^9, 2^{-11}, 2^2)$

4%, FAU AIBO was reduced 7%, and EMO-DB was most affected with 26% reduction.

For these small feature sets, FSVM shows better performance than SVM. In EMO-DB, FSVM improves 4%. In FAU AIBO, FSVM shows a slight gain in UA but 4% on WA measure. However improvement is only minor in aGender. In ANDOSL, FSVM does not show any gain.

Table 4.9: Classification rates (%) of SVM and FSVM on the four data sets with SFFS feature selection.

	SVM+SFFS			FSVM+SFFS		
	UA	WA	Parameters (C, γ)	UA	WA	Parameters (C, γ, β)
ANDOSL	98.24	98.24	$(2^3, 2^{-1})$	98.25	98.25	$(2^{15}, 2^{-1}, 2^0)$
aGender	41.19	42.18	$(2^1, 2^{-3})$	41.2	42.18	$(2^1, 2^{-3}, 2^1)$
FAU AIBO	38.18	56.47	$(2^9, 2^{-1})$	38.21	60.19	$(2^{11}, 2^{-3}, 2^{-1})$
EMO-DB	55.82	59.72	$(2^7, 2^{-15})$	57.02	60.73	$(2^7, 2^{-15}, 2^2)$

For small data sets, such as EMO-DB with only 384 vectors, SFFS causes performance reduction. However, for large data sets SFFS helps reduce large amount of training and testing time while performance slightly decreases as in ANDOSL and aGender data sets. Since Correlation-based Feature Subset Selection is an automatic

and unsupervised feature selection method based on statistical properties of the data, the resulting features are generally not the best feature set for every classifier or the most representative sets that encode paralinguistic information [55].

Table 4.10 shows the detailed classification rates of each classes within each data set. Performance of SVM and FSVM, with and without SFFS, are shown. The abbreviations are explained in Section 4.1. The classification results are high in ANDOSL data set on all four data sets with SFFS feature selection. In aGender data set, Adult Male (AF) was hardly recognised, 30.5%, but was slighly increased to 33.7% after feature selection. However the Youth Male (YM) classification rate was reduced significantly after feature selection, from 42.9% to 24.2%. In FAU AIBO data set, the Rest (R) emotion was almost unrecognisable and was even worse after feature selection. FSVM reduced the classification rate 4% of the Positive (P) emotion but gained nearly 8% of the Neutral (N) emotion. In EMO-DB, FSVM gains nearly 5% of Happiness (F), both with and without SFFS, and 2% of sadness (T) with SFFS.

4.4 Feature Type Relevance for Age and Gender Classification

This experiment determines which features in the speech signal are suited to represent age and gender in human speech. Experiments were extensively performed in aGender and ANDOSL data sets to automatically discriminate between age groups. The 1582 acoustic features were extracted in three steps: first, the 38 low-level descriptors (see Table 4.11); next, their first order regression coefficients were added; then, 21 functionals (see Table 4.11) were applied. These 1580 acoustic features were grouped into six groups:

1. MFCCs [0-14]
2. Log Mel Frequency Band [0-7]
3. LSP Frequency [0-7]

Table 4.10: Classification rates of SVM and FSVM on the four data sets

		SVM	FSVM	SVM+SFFS	FSVM+SFFS
ANDOSL	FE	99.17	99.17	98.19	98.22
	FM	99.67	99.67	97.67	97.67
	FY	98.58	98.58	98.47	98.5
	ME	99.33	99.33	98.28	98.28
	MM	99.08	99.08	98.25	98.22
	MY	99.36	99.36	98.58	98.61
aGender	CX	52.92	52.92	48.31	48.6
	YF	51.32	51.32	52.9	52.65
	YM	42.9	42.95	24.19	24.24
	AF	39.6	39.63	30.82	30.79
	AM	30.45	30.53	33.68	33.72
	SF	44.09	44.12	38.61	38.58
	SM	54.34	54.29	59.8	59.8
FAU AIBO	A	39.84	40.41	32.35	33.03
	E	45.48	45.2	40.04	39.46
	N	82.42	82.4	75.19	82.86
	P	46.29	46.29	30.42	26.41
	R	15.4	15.4	12.9	9.29
EMO-DB	W	93.7	90.55	73.23	73.23
	L	84.81	84.81	70.89	70.89
	E	81.58	78.95	34.21	34.21
	A	78.18	78.18	43.64	45.45
	F	64.06	68.75	40.63	45.31
	T	81.13	81.13	67.92	69.81
	N	89.74	91.03	60.26	60.26

4. PCM loudness

5. Pitch related (F_0, F_0 Envelope, and Voicing Probability)

6. Jitter and Shimmer (Jitter local, Jitter consecutive frame pairs, Shimmer local)

Finally, the two single features F_0 number of onsets and turn duration were added to each of these six groups. The first three groups above are spectral features, the next two groups are prosodic features, and the final group are the voice quality features. Correlation based feature subset selection with SFFS was employed to eliminate redundant features and to improve performance of classifiers because of large data sets. The lists of remaining features after SFFS for each feature groups are shown in section 5.2. For each of these groups, classification results using SVM and FSVM were reported.

Table 4.11: 38 low-level descriptors with regression coefficients and 21 functionals.

Descriptors	**Functionals**
PCM loudness	Position max./min.
MFCCs [0-14]	arithmetic mean, std. deviation
log Mel Frequency Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	linear regression coefficients 1/2
F_0	linear regression error Q/A
F_0 Envelope	quartile 1/2/3
Voicing Probability	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter consecutive frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

Experiments were carried out in 3-fold cross validation manner in ANDOSL data set, except for the aGender data set since it had separated training and developing sets already. At first, the data set was separated into three folds. Next, a fold

was considered as the validation set and the rest were the training set. The cross validation accuracy was the average of accuracies on predicting the validation sets [12]. Parameter selection procedures were employed to find the best parameters as follows: a possible interval of C and γ for SVM or additional β for FSVM were chosen as the grid space, $C = 2^1, 2^3, \ldots, 2^{15}$, $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^3$, $\beta = 2^{-2}, 2^{-1}, \ldots, 2^3$. Then, all grid points of (C, γ) for SVM or (C, γ, β) for FSVM were tested to find which one gave the highest cross validation accuracy. The best parameter set was then selected. For evaluation measures we used unweighted accuracy (UA) as the main measure to better reflect unbalance among classes and weighted accuracy (WA) as an extra measure.

To evaluate which feature types are essential for detecting specific age and gender groups, binary classification experiments were conducted for every possible class pair in each data set using only one of the six feature type sets described above. These experiments were performed in aGender and ANDOSL data sets.

Tables 4.12, 4.13, 4.14, 4.15, 4.16 and 4.17 show the classification result of each pair of only Age on aGender and ANDOSL data sets employing SVM and FSVM respectively. As can be seen, the classification rate of consecutive age groups are lower than distant age groups, which is similar to human perception. The all feature set showed the best performance. The MFCCs features show highest accuracies among feature groups for both data sets. The second highest group is log mel frequency in ANDOSL and F_0 in aGender data set. PCM loudness shows lowest performance. It hardly differentiated between Elderly and Young in ANDOSL data set (71%) or between Child and Senior in aGender (66%).

Table 4.12: Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM (ANDOSL data set).

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
FE/FM	98.57	0.57	-9.4	-11.35	-28.42	-24.26	-36.25
FE/FY	99.06	0.18	-10.65	-11.43	-34.35	-17.18	-32.44
FE/ME	99.93	0.04	-3.76	-3.6	-29.86	-2.1	-12.54
FE/MM	99.93	0.04	-2.81	-4.19	-20.19	-2.54	-9.25
FE/MY	99.92	0.03	-3.22	-2.11	-25.44	-1.24	-9.5
FM/FY	98.29	0.76	-14.01	-19.26	-31.64	-21.97	-32.44
FM/MM	99.96	0.03	-2.97	-3.29	-16.47	-2.22	-11.26
FM/MY	99.99	-0.03	-3.39	-3.17	-19.78	-1.4	-11.5
FY/MY	99.99	0	-4.11	-1.82	-26	-0.11	-7.33
ME/FM	99.99	0.01	-2.31	-1.82	-14.43	-2.39	-14.42
ME/FY	100	0	-3.35	-1.5	-26.24	-0.96	-8.19
ME/MM	99.07	0.35	-7.36	-5.51	-28.56	-30.19	-34.08
ME/MY	98.88	0.07	-7.17	-5.9	-27.28	-26.56	-32.67
MM/FY	100	-0.01	-3.24	-2.79	-24.38	-0.14	-6.29
MM/MY	98.94	0.33	-8.07	-9.35	-27.82	-32.07	-41.39

Table 4.13: Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM (aGender data set).

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
CX/YF	68.84	2.32	-10.33	-12.27	-10.86	1.56	-10.39
CX/YM	91.02	-1.58	-19.05	-21.33	-19.09	-0.74	-5.1
CX/AF	79.11	-2.81	-14.12	-24.75	-16.71	-1.34	-18.48
CX/AM	92.34	-1.61	-19.86	-22.02	-19.62	-1.2	-7.13
CX/SF	77.96	-0.97	-17.37	-20.07	-16.77	2.01	-14.57
CX/SM	93.28	-1.54	-19.35	-21.4	-22.1	-0.93	-7.05
YF/YM	95.85	-2.84	-20.52	-20.36	-20.24	-0.29	-8.65
YF/AF	69.19	-3.5	-13.97	-15.3	-14.48	-1.74	-15.04
YF/AM	96.66	-3.76	-21.76	-19.45	-21.42	-0.86	-10.03
YF/SF	72.61	-4.62	-17.14	-15.55	-18.4	0.68	-12.84
YF/SM	96.64	-3.28	-23.49	-19.36	-23.53	-1.54	-9.68
YM/AF	93.53	-4.5	-17.18	-20.36	-15.82	-0.24	-8.62
YM/AM	57.9	-4.96	-5.32	-2.58	-5.98	-4.4	-5.72
YM/SF	91.08	-4.12	-14.67	-18.55	-13.89	-2.22	-7.87
YM/SM	65.19	0.27	-4.85	-1.02	-3.5	-3.74	-2.33
AF/AM	94.19	-5.19	-18.47	-18.42	-15.73	-1.21	-10.81
AF/SF	55.74	2.47	0.82	-4.28	1.79	-1.76	-4.59
AF/SM	93.46	-5.11	-19.91	-20.06	-18.38	-2.09	-9.81
AM/SF	92.29	-4.81	-15.63	-17.47	-15.07	-3.56	-10.39
AM/SM	61.68	-0.85	-4.24	-1.51	-0.77	-6.6	-8.85
SF/SM	90.16	-3.98	-18.38	-18.83	-15.34	-2.38	-8.88

Table 4.14: Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM (ANDOSL data set)

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
FE/FM	98.57	0.57	-9.39	-11.33	-28.42	-25.24	-35.72
FE/FY	99.06	0.18	-10.67	-11.43	-34.35	-17.9	-31.89
FE/ME	99.93	0.04	-3.82	-3.57	-29.88	-2.07	-12.49
FE/MM	99.93	0.03	-2.82	-4.17	-20.21	-2.92	-9.32
FE/MY	99.92	0.03	-3.24	-2.13	-25.47	-2.88	-10.04
FM/FY	98.29	0.76	-14.01	-19.26	-31.64	-22.18	-32.68
FM/MM	99.97	0	-2.97	-3.33	-16.53	-2.49	-11.26
FM/MY	99.99	-0.03	-3.38	-3.17	-19.76	-1.72	-11.54
FY/MY	99.97	0.01	-4.14	-1.83	-26.01	-0.19	-7.44
ME/FM	100	-0.06	-2.35	-1.82	-14.44	-2.5	-14.32
ME/FY	100	0	-3.33	-1.54	-26.19	-1.13	-8.28
ME/MM	99.07	0.35	-7.38	-5.6	-28.56	-30.74	-34.03
ME/MY	98.88	0.07	-7.17	-5.9	-27.29	-28.44	-33.18
MM/FY	100	-0.01	-3.36	-2.88	-24.35	-0.43	-6.33
MM/MY	98.94	0.33	-8.07	-9.36	-27.86	-32.33	-41.69

Table 4.15: Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM (aGender data set)

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
CX/YF	68.84	2.32	-10.33	-15.04	-10.86	1.56	-10.39
CX/YM	91	-1.69	-19.03	-20.87	-19.07	-0.68	-5.12
CX/AF	79.11	-2.81	-14.12	-25.55	-16.71	-1.34	-18.48
CX/AM	92.34	-1.75	-19.86	-19.41	-19.62	-1.2	-7.17
CX/SF	77.96	-0.97	-17.37	-21.16	-16.77	1.98	-14.57
CX/SM	93.17	-1.45	-19.24	-18.34	-21.98	-0.74	-6.91
YF/YM	95.85	-3.11	-20.52	-18.83	-20.24	-0.37	-8.57
YF/AF	69.19	-3.5	-13.97	-15.75	-14.48	-1.74	-14.91
YF/AM	96.64	-3.84	-21.74	-16.22	-21.4	-0.86	-10.18
YF/SF	72.61	-4.62	-17.14	-15.88	-18.4	0.68	-12.76
YF/SM	96.58	-3.27	-23.43	-16.66	-23.47	-1.5	-9.79
YM/AF	93.67	-4.68	-17.32	-18.93	-15.96	-0.31	-8.8
YM/AM	57.9	-4.96	-5.32	-2.93	-5.98	-4.4	-5.7
YM/SF	91.07	-4.08	-14.66	-15.91	-13.87	-2.23	-7.83
YM/SM	65.19	0.27	-4.85	-2.92	-3.5	-3.74	-2.33
AF/AM	94.19	-5.19	-18.47	-16.38	-15.73	-1.21	-10.81
AF/SF	55.74	2.47	0.82	-3.44	1.79	-1.76	-4.59
AF/SM	93.39	-5.02	-19.84	-16.71	-18.31	-2.06	-9.75
AM/SF	92.28	-4.79	-15.61	-14.59	-15.05	-3.44	-10.42
AM/SM	61.68	-0.85	-4.24	-3.93	-0.77	-6.6	-8.85
SF/SM	90.16	-4.07	-18.38	-14.67	-15.34	-2.57	-8.81

Table 4.16: Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM. Averaging from Table 4.12 and Table 4.13

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
Elderly/Middle	99.39	0.24	-5.47	-5.72	-22.9	-14.85	-23.5
Elderly/Young	99.46	0.07	-6.1	-5.24	-28.33	-11.48	-20.7
Middle/Young	99.31	0.26	-7.18	-8.64	-25.9	-13.9	-22.91
Child/Youth	79.93	0.37	-14.69	-16.8	-14.98	0.41	-7.75
Child/Adult	85.72	-2.21	-16.99	-23.38	-18.16	-1.27	-12.8
Child/Senior	85.62	-1.26	-18.36	-20.74	-19.43	0.54	-10.81
Youth/Adult	79.32	-4.18	-14.56	-14.42	-14.43	-1.81	-9.86
Youth/Senior	81.38	-2.94	-15.04	-13.62	-14.83	-1.7	-8.18
Adult/Senior	75.79	-2.07	-9.74	-10.83	-8.11	-3.5	-8.41

Table 4.17: Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM. Averaging from Table 4.14 and Table 4.15

	All features	MFCCs	log Mel Freq.	LSP Freq.	PCM loudness	F_0 features	Jitter & Shimmer
Elderly/Middle	99.39	0.22	-5.48	-5.73	-22.91	-15.35	-23.35
Elderly/Young	99.46	0.07	-6.1	-5.25	-28.33	-12.59	-20.85
Middle/Young	99.31	0.26	-7.2	-8.67	-25.9	-14.17	-23.06
Child/Youth	79.92	0.32	-14.68	-17.95	-14.97	0.44	-7.76
Child/Adult	85.72	-2.28	-16.99	-22.48	-18.16	-1.27	-12.82
Child/Senior	85.57	-1.21	-18.3	-19.75	-19.38	0.62	-10.74
Youth/Adult	79.35	-4.24	-14.59	-13.46	-14.46	-1.83	-9.9
Youth/Senior	81.36	-2.93	-15.02	-12.84	-14.81	-1.7	-8.18
Adult/Senior	75.77	-2.05	-9.72	-9.67	-8.09	-3.46	-8.41

Chapter 5

Conclusions and Future Research

5.1 Conclusions

I have presented speaker classification based on the three Australian accents which are Broad, General and Cultivated, using the ANDOSL database consisting of 108 speakers, each speaking 200 long utterances. I used GMMs with different Gaussian components and SVM with RBF kernel as classification methods. For GMMs, I considered the classification rates versus the number of Gaussian components, age, and gender. I extracted MFCCs features for speech and used those features to train Gaussian speaker models. I extracted 16 useful low-level descriptors: zero-crossing-rate (ZCR), root mean square (RMS) frame energy, pitch frequency, harmonics-to-noise ratio (HNR), and mel-frequency cepstral coefficients (MFCCs) and applied the 12 functionals mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE) to train SVM. The one-against-one strategy was used for multi-class classification problems resulted in $n(n-1)/2$ class pairs SVM model where n is the total number of classes. Most classification rates were high, ranging from 90% to 99%. The Cultivated accent was hard to recognise for elderly female speakers.

I have also presented a fuzzy approach to support vector machine and have applied this method to age and gender classification. Experimental results on the aGender,

ANDOSL, FAU AIBO, and EMO-DB databases have shown that the proposed fuzzy support vector machine could improve the classification accuracy for age, gender, and emotion classification, FSVM showed improvement on aGender compared with the baseline results for the INTERSPEECH 2010 Paralinguistic Challenge. The features chosen for age and gender classification task include 29 LLD of Mel-Frequency cepstral coefficients 0-14, 8 line spectral pair frequencies computed from 8 LPC coefficients, F_0, F_0 Envelope, Voicing Probability, Jitter local, Jitter consecutive frame pairs, shimmer local and applied 8 functionals include arithmetic mean, standard deviation, skewness, kurtosis, percentile 1/99,percentile range 99-1. Experiments were run in 5-fold Stratified-Cross-Validation manner for all databases except for the experiment on aGender database with separated train and develop sets.

I have performed experiments to determine which features in the speech signal are suited to represent age and gender in human speech, several experiments were performed to automatically discriminate between varieties of age groups. The 1582 acoustic features and transliteration were extracted in three steps: first, the 38 low-level descriptors include PCM loudness, MFCCs [0-14], log Mel Freq. Band [0-7], LSP Frequency [0-7], F_0, F_0 Envelope, Voicing Probability, Jitter local, Jitter consecutive frame pairs, Shimmer local are extracted and smoothed by simple moving average low-pass filtering. Next, their first order regression coefficients were added in full HTK compliance. Then, 21 functionals, including position max/min, arithmetic mean, standard deviation, skewness, kurtosis, linear regression coefficient 1/2, linear regression error Q/A, quartile 1/2/3, quartile range 2-1/3-2/3-1, percentile 1/99, percentile range 99-1, up-level time 75/90, were applied. 16 zero-information features were discarded and the 2 single features F_0 number of onsets and turn duration are added. These 1582 acoustic features were grouped into six groups including PCM loudness, MFCCs [0-14], log Mel Frequency Band [0-7], LSP Frequency [0-7], pitch related (F_0, F_0 Envelope, and Voicing Probability), Jitter and Shimmer (Jitter local, Jitter consecutive frame pairs, Shimmer local). Correlation based feature subset selection with SFFS was employed to eliminate redundant features because of large feature sets. The experiments proved that spectral features contain the most relevant

information about age and gender within speech for almost every pair of age and gender for both databases. When using only LSP features for age and gender recognition, performance was shown to be 6.9% higher compared to the average. Cepstral features performed even 7.1% better than the average feature type. Pitch, as a prosodic Low-Level-Descriptor, prevailed only for the pair male/female where it performed 6.3% better than the average.

5.2 Future Research

For my future research, I will perform experiments on Australian foreign accent data sets using new versions of SVM for classification such as multi-sphere SVM and multi-sphere support vector data description. Various high-level features will be employed. Discussion and comparison to these findings would be carried out in order to determine good feature set in combination with new techniques.

Classification between male, female and children speaker is still a hard problem [52]. Future research would explore current successful methods as presented in [36]. In this work, the gender classification system was composed by the fusion of several sub-systems Gaussian mixture models-Universal background model (GMM-UBM), Multi-Layer Perceptrons (MLP) and SVM. This system also used additional speech data sets for training models. The purpose was to handle more speaker variability as well as diverse audio background conditions. This increased variability may help to increase the robustness of gender detection system. In addition, the final results were obtained by the calibration and linear logistic regression fusion back-end. In some applications such as the automatic detection of child abuse videos on the web, the detection of children's voices is important.

Classification of speaker age is even more difficult [52]. Future research would explore current effective methods as presented in [33]. In this work, acoustic frame-based feature sets, as well as utterance-based acoustic, prosodic and voice quality features were extracted. The models included GMMs and SVM with linear Gaussian backends and logistic regression-based fusion. Maximum-Mutual-Information (MMI)-

training and Joint Factor-Analysis (JFA) were applied.

Appendices

Remaining features of ANDOSL, EMO-DB, FAU AIBO and aGender data sets after SFFS. Information Gain Ratio is shown in the left of each data set.

ANDOSL (450 → 56)

0.04364	mfcc_sma[0]_linregerrQ
0.06312	mfcc_sma[0]_percentile1.0
0.02150	mfcc_sma[1]_kurtosis
0.09726	mfcc_sma[3]_percentile1.0
0.04740	mfcc_sma[3]_percentile99.0
0.07025	mfcc_sma[4]_amean
0.05023	mfcc_sma[5]_amean
0.08968	mfcc_sma[6]_amean
0.04140	mfcc_sma[7]_amean
0.07446	mfcc_sma[9]_amean
0.01111	mfcc_sma[10]_kurtosis
0.02670	mfcc_sma[10]_percentile99.0
0.17968	mfcc_sma[11]_amean
0.15975	mfcc_sma[11]_skewness
0.17596	mfcc_sma[11]_percentile1.0
0.13702	mfcc_sma[11]_percentile99.0
0.03654	mfcc_sma[12]_amean
0.12797	mfcc_sma[13]_amean
0.12255	mfcc_sma[13]_percentile99.0
0.20975	mfcc_sma[14]_amean
0.13457	mfcc_sma[14]_linregerrQ
0.20589	mfcc_sma[14]_percentile1.0
0.14288	mfcc_sma[14]_percentile99.0
0.04284	lspFreq_sma[0]_percentile1.0
0.03163	lspFreq_sma[1]_amean
0.07332	lspFreq_sma[1]_stddev
0.08807	lspFreq_sma[3]_skewness
0.06085	lspFreq_sma[5]_stddev
0.11472	lspFreq_sma[7]_amean
0.24179	F0finEnv_sma_amean
0.22946	F0finEnv_sma_linregerrQ
0.22036	F0finEnv_sma_stddev
0.20142	voicingFinalUnclipped_sma_percentile99.0
0.01016	mfcc_sma_de[0]_skewness
0.00915	mfcc_sma_de[1]_skewness
0.01287	mfcc_sma_de[2]_skewness
0.01443	mfcc_sma_de[4]_amean
0.01757	mfcc_sma_de[4]_skewness
0.00543	mfcc_sma_de[5]_skewness
0.00887	mfcc_sma_de[6]_skewness
0.00208	mfcc_sma_de[8]_skewness
0.00476	mfcc_sma_de[9]_skewness
0.00935	mfcc_sma_de[10]_skewness
0.02548	mfcc_sma_de[11]_skewness
0.01307	mfcc_sma_de[12]_skewness
0.00541	mfcc_sma_de[13]_skewness
0.00964	mfcc_sma_de[14]_skewness
0.04315	lspFreq_sma_de[0]_kurtosis
0.02030	lspFreq_sma_de[2]_kurtosis
0.03461	lspFreq_sma_de[6]_kurtosis
0.19299	F0finEnv_sma_de_stddev
0.06042	voicingFinalUnclipped_sma_de_skewness
0.26651	F0final_sma_amean
0.12722	jitterDDP_sma_amean
0.18978	F0final_sma_de_stddev
0.22813	F0final_sma_de_percentile99.0

EMO-DB (384 → 38)

0.3918	pcm_RMSenergy_sma_min
0.1833	pcm_RMSenergy_sma_range
0.2284	pcm_RMSenergy_sma_linregc1

0.3058	mfcc_sma[1]_amean
0.2002	mfcc_sma[1]_skewness
0.3525	mfcc_sma[2]_min
0.4142	mfcc_sma[2]_amean
0.3410	mfcc_sma[2]_linregerrQ
0.2709	mfcc_sma[4]_linregc2
0.3338	mfcc_sma[5]_amean
0.3115	mfcc_sma[5]_linregerrQ
0.1432	mfcc_sma[6]_amean
0.0819	mfcc_sma[8]_skewness
0.1994	mfcc_sma[9]_amean
0.0962	mfcc_sma[10]_max
0.1526	mfcc_sma[11]_linregc1
0.3574	pcm_zcr_sma_amean
0.0958	voiceProb_sma_minPos
0.1263	voiceProb_sma_linregc1
0.2346	F0_sma_maxPos
0.2062	pcm_RMSenergy_sma_de_amean
0.1966	mfcc_sma_de[1]_linregc2
0.1625	mfcc_sma_de[1]_linregerrQ
0.3845	mfcc_sma_de[2]_linregerrQ
0.1938	mfcc_sma_de[3]_linregc1
0.1182	mfcc_sma_de[3]_linregerrQ
0.3605	mfcc_sma_de[5]_linregerrQ
0.0989	mfcc_sma_de[6]_kurtosis
0.1158	mfcc_sma_de[7]_linregc2
0.1915	mfcc_sma_de[8]_max
0.1858	mfcc_sma_de[8]_linregc1
0.2992	mfcc_sma_de[9]_stddev
0.2425	mfcc_sma_de[10]_stddev
0.3335	mfcc_sma_de[12]_linregerrQ
0.3373	mfcc_sma_de[12]_stddev
0.1295	voiceProb_sma_de_skewness
0.1735	voiceProb_sma_de_kurtosis
0.3418	F0_sma_de_stddev

FAU AIBO (384 → 31)

0.06046	pcm_RMSenergy_sma_range
0.05437	pcm_RMSenergy_sma_amean
0.01803	pcm_RMSenergy_sma_linregc1
0.04188	pcm_RMSenergy_sma_linregc2
0.06285	pcm_RMSenergy_sma_linregerrQ
0.06114	pcm_RMSenergy_sma_stddev
0.02125	mfcc_sma[1]_max
0.03679	mfcc_sma[1]_amean
0.02510	mfcc_sma[1]_linregc2
0.02855	mfcc_sma[4]_min
0.02764	mfcc_sma[4]_linregerrQ
0.02766	mfcc_sma[4]_stddev
0.01985	voiceProb_sma_skewness
0.02045	F0_sma_kurtosis
0.05547	pcm_RMSenergy_sma_de_max
0.04483	pcm_RMSenergy_sma_de_min
0.03836	pcm_RMSenergy_sma_de_linregc1
0.05164	pcm_RMSenergy_sma_de_linregc2
0.05694	pcm_RMSenergy_sma_de_linregerrQ
0.02943	mfcc_sma_de[1]_max
0.02359	mfcc_sma_de[1]_stddev
0.02101	mfcc_sma_de[2]_min
0.02503	mfcc_sma_de[2]_skewness
0.02639	mfcc_sma_de[2]_kurtosis
0.02959	mfcc_sma_de[4]_range
0.00825	mfcc_sma_de[5]_skewness
0.04188	pcm_zcr_sma_de_min
0.03012	pcm_zcr_sma_de_stddev
0.02502	F0_sma_de_max
0.02961	F0_sma_de_min
0.02539	F0_sma_de_range

aGender (450 → 23)

0.013779	mfcc_sma[3]_stddev
0.020062	mfcc_sma[5]_percentile1.0
0.051679	mfcc_sma[8]_percentile1.0
0.014573	mfcc_sma[9]_amean
0.048035	mfcc_sma[10]_amean
0.076544	mfcc_sma[13]_percentile1.0
0.088656	mfcc_sma[14]_stddev
0.014226	lspFreq_sma[4]_kurtosis
0.025966	lspFreq_sma[4]_percentile1.0
0.025172	lspFreq_sma[6]_pctlrange0-1
0.135367	F0finEnv_sma_amean
0.168868	voicingFinalUnclipped_sma_percentile99.0
0.005366	mfcc_sma_de[6]_skewness
0.006261	mfcc_sma_de[7]_skewness
0.007759	lspFreq_sma_de[0]_kurtosis
0.003741	lspFreq_sma_de[2]_skewness
0.119601	F0finEnv_sma_de_linregerrQ
0.111638	F0finEnv_sma_de_percentile1.0
0.016999	voicingFinalUnclipped_sma_de_stddev
0.156454	F0final_sma_amean
0.099535	F0final_sma_skewness
0.129106	jitterDDP_sma_amean
0.128105	F0final_sma_de_percentile99.0

Remaining features of each feature group after SFFS of aGender data set. Information Gain Ratio is shown in the left of each feature group.

MFCC [0-14] (632 → 34)

0.018771	mfcc_sma[0]_stddev
0.022554	mfcc_sma[2]_percentile1.0
0.015963	mfcc_sma[3]_linregc2
0.023987	mfcc_sma[5]_linregc2
0.011643	mfcc_sma[6]_quartile1
0.046795	mfcc_sma[8]_linregc2
0.051435	mfcc_sma[8]_percentile1.0
0.014272	mfcc_sma[9]_amean
0.046795	mfcc_sma[10]_amean
0.053380	mfcc_sma[10]_linregc1
0.065988	mfcc_sma[10]_linregc2
0.060298	mfcc_sma[10]_percentile1.0
0.046310	mfcc_sma[11]_percentile1.0
0.060368	mfcc_sma[12]_stddev
0.038828	mfcc_sma[12]_percentile1.0
0.037593	mfcc_sma[12]_percentile99.0
0.049803	mfcc_sma[13]_linregc2
0.072219	mfcc_sma[13]_stddev
0.038363	mfcc_sma[13]_skewness
0.076944	mfcc_sma[13]_percentile1.0
0.074193	mfcc_sma[13]_pctlrange0-1
0.082804	mfcc_sma[14]_linregerrA
0.082403	mfcc_sma[14]_stddev
0.039498	mfcc_sma[14]_skewness
0.042056	mfcc_sma[14]_percentile1.0
0.077973	mfcc_sma[14]_percentile99.0
0.088575	mfcc_sma[14]_pctlrange0-1
0.005442	mfcc_sma_de[6]_skewness
0.055453	mfcc_sma_de[13]_pctlrange0-1
0.098101	mfcc_sma_de[14]_amean
0.044293	mfcc_sma_de[14]_kurtosis
0.059306	mfcc_sma_de[14]_percentile1.0
0.073667	mfcc_sma_de[14]_pctlrange0-1
0.106454	F0final___Turn_numOnsets

Log Mel Frequency Band [0-7] (338→4)

0.02830	logMelFreqBand_sma[3]_percentile99.0
0.00672	logMelFreqBand_sma[7]_percentile1.0
0.00767	logMelFreqBand_sma_de[0]_quartile2
0.10645	F0final___Turn_numOnsets

LSP Frequency [0-7] (338→6)

0.014344	lspFreq_sma[4]_kurtosis
0.026129	lspFreq_sma[4]_percentile1.0
0.027516	lspFreq_sma[6]_percentile1.0
0.014090	lspFreq_sma[7]_percentile99.0
0.003807	lspFreq_sma_de[2]_skewness
0.106454	F0final___Turn_numOnsets

PCM loudness (44→3)

0.020794	pcm_loudness_sma_pctlrange0-1
0.002474	pcm_loudness_sma_de_skewness
0.106454	F0final___Turn_numOnsets

F0 features (124→11)

0.13536	F0finEnv_sma_amean
0.14547	F0finEnv_sma_quartile1
0.16894	voicingFinalUnclipped_sma_percentile99.0
0.11931	F0finEnv_sma_de_stddev
0.11166	F0finEnv_sma_de_percentile1.0
0.01700	voicingFinalUnclipped_sma_de_stddev
0.11114	F0final_sma_minPos
0.14311	F0final_sma_linregc2
0.15365	F0final_sma_quartile1
0.16947	F0final_sma_quartile2
0.15324	F0final_sma_quartile3

Jitter and Shimmer (116→22)

0.10458	jitterLocal_sma_skewness
0.10993	jitterLocal_sma_upleveltime90
0.13099	jitterDDP_sma_amean
0.12237	jitterDDP_sma_linregerrA
0.11905	jitterDDP_sma_linregerrQ
0.11288	jitterDDP_sma_kurtosis
0.10310	jitterDDP_sma_quartile1
0.10933	jitterDDP_sma_quartile2
0.11392	jitterDDP_sma_quartile3
0.11474	jitterDDP_sma_iqr2-3
0.11781	jitterDDP_sma_percentile99.0
0.11049	jitterDDP_sma_upleveltime75
0.12347	jitterDDP_sma_upleveltime90
0.11433	shimmerLocal_sma_stddev
0.09780	shimmerLocal_sma_skewness
0.10855	shimmerLocal_sma_kurtosis
0.10821	shimmerLocal_sma_upleveltime90
0.12196	jitterDDP_sma_de_skewness
0.11005	jitterDDP_sma_de_quartile1

0.11929	jitterDDP_sma_de_percentile99.0
0.11718	jitterDDP_sma_de_upleveltime90
0.10234	shimmerLocal_sma_de_upleveltime75
All feature (1582→25)	
0.013776	mfcc_sma[3]_stddev
0.024024	mfcc_sma[5]_linregc2
0.051472	mfcc_sma[8]_percentile1.0
0.013472	mfcc_sma[9]_linregc2
0.069241	mfcc_sma[10]_linregc2
0.018075	mfcc_sma[11]_quartile2
0.024362	logMelFreqBand_sma[7]_stddev
0.014358	lspFreq_sma[4]_kurtosis
0.025908	lspFreq_sma[4]_percentile1.0
0.000000	F0finEnv_sma_minPos
0.087542	F0finEnv_sma_kurtosis
0.003710	voicingFinalUnclipped_sma_percentile1.0
0.005385	mfcc_sma_de[6]_skewness
0.004187	mfcc_sma_de[7]_upleveltime75
0.007710	logMelFreqBand_sma_de[0]_quartile2
0.007820	lspFreq_sma_de[0]_kurtosis
0.102646	F0finEnv_sma_de_linregc2
0.071988	F0finEnv_sma_de_iqr2-3
0.017074	voicingFinalUnclipped_sma_de_stddev
0.085861	F0final_sma_maxPos
0.096300	F0final_sma_linregc1
0.101255	F0final_sma_kurtosis
0.101255	F0final_sma_kurtosis
0.101255	F0final_sma_kurtosis
0.093851	jitterDDP_sma_minPos

Remaining features of each feature group after SFFS of ANDOSL data set. Information Gain Ratio is shown in the left of each feature group.

MFCC [0-14] (632 →68)	
0.04364	mfcc_sma[0]_linregerrQ
0.06312	mfcc_sma[0]_percentile1.0
0.02531	mfcc_sma[0]_upleveltime75
0.04979	mfcc_sma[1]_quartile3
0.01537	mfcc_sma[1]_percentile1.0
0.02744	mfcc_sma[2]_pctlrange0-1
0.09726	mfcc_sma[3]_percentile1.0
0.04740	mfcc_sma[3]_percentile99.0
0.07025	mfcc_sma[4]_amean
0.06009	mfcc_sma[5]_quartile1
0.02595	mfcc_sma[6]_minPos
0.08968	mfcc_sma[6]_amean
0.08582	mfcc_sma[6]_percentile1.0
0.04140	mfcc_sma[7]_amean
0.04516	mfcc_sma[8]_amean
0.03258	mfcc_sma[9]_linregerrA
0.08453	mfcc_sma[9]_quartile3
0.01415	mfcc_sma[9]_upleveltime90
0.07818	mfcc_sma[10]_amean
0.08893	mfcc_sma[10]_quartile2
0.04579	mfcc_sma[11]_minPos
0.17968	mfcc_sma[11]_amean
0.10379	mfcc_sma[11]_linregc2
0.15975	mfcc_sma[11]_skewness
0.17976	mfcc_sma[11]_quartile1
0.12959	mfcc_sma[11]_iqr1-2
0.17596	mfcc_sma[11]_percentile1.0
0.13702	mfcc_sma[11]_percentile99.0
0.08648	mfcc_sma[11]_upleveltime75
0.05275	mfcc_sma[12]_linregc1
0.02626	mfcc_sma[12]_quartile3
0.02522	mfcc_sma[12]_pctlrange0-1
0.02535	mfcc_sma[13]_minPos
0.12797	mfcc_sma[13]_amean
0.09462	mfcc_sma[13]_linregc2
0.13081	mfcc_sma[13]_quartile3
0.12255	mfcc_sma[13]_percentile99.0
0.20975	mfcc_sma[14]_amean
0.14624	mfcc_sma[14]_linregc2
0.13457	mfcc_sma[14]_linregerrQ
0.21614	mfcc_sma[14]_quartile1
0.20331	mfcc_sma[14]_quartile2
0.19106	mfcc_sma[14]_quartile3
0.16419	mfcc_sma[14]_iqr1-2
0.20589	mfcc_sma[14]_percentile1.0
0.14288	mfcc_sma[14]_percentile99.0
0.09061	mfcc_sma[14]_upleveltime75
0.01016	mfcc_sma_de[0]_skewness
0.02301	mfcc_sma_de[1]_kurtosis

0.01363	mfcc_sma_de[1]_quartile2
0.01287	mfcc_sma_de[2]_skewness
0.00525	mfcc_sma_de[3]_quartile2
0.01443	mfcc_sma_de[4]_amean
0.01757	mfcc_sma_de[4]_skewness
0.00543	mfcc_sma_de[5]_skewness
0.00887	mfcc_sma_de[6]_skewness
0.00536	mfcc_sma_de[7]_upleveltime75
0.00208	mfcc_sma_de[8]_skewness
0.00476	mfcc_sma_de[9]_skewness
0.00935	mfcc_sma_de[10]_skewness
0.10429	mfcc_sma_de[11]_linregc1
0.02548	mfcc_sma_de[11]_skewness
0.01307	mfcc_sma_de[12]_skewness
0.00630	mfcc_sma_de[13]_quartile2
0.02242	mfcc_sma_de[13]_pctlrange0-1
0.08816	mfcc_sma_de[14]_linregc1
0.10399	mfcc_sma_de[14]_pctlrange0-1
0.05487	F0final___Turn_numOnsets
Log Mel Frequency Band [0-7] (338→25)	
0.06160	logMelFreqBand_sma[0]_percentile1.0
0.06789	logMelFreqBand_sma[0]_pctlrange0-1
0.01721	logMelFreqBand_sma[1]_maxPos
0.05346	logMelFreqBand_sma[1]_pctlrange0-1
0.04919	logMelFreqBand_sma[3]_linregerrA
0.03188	logMelFreqBand_sma[3]_kurtosis
0.06035	logMelFreqBand_sma[3]_upleveltime75
0.03473	logMelFreqBand_sma[3]_upleveltime90
0.04214	logMelFreqBand_sma[5]_iqr1-2
0.06556	logMelFreqBand_sma[5]_percentile1.0
0.04581	logMelFreqBand_sma[5]_pctlrange0-1
0.03099	logMelFreqBand_sma[5]_upleveltime75
0.05211	logMelFreqBand_sma[6]_linregerrQ
0.05492	logMelFreqBand_sma[6]_stddev
0.06473	logMelFreqBand_sma[6]_percentile1.0
0.04181	logMelFreqBand_sma[6]_pctlrange0-1
0.03433	logMelFreqBand_sma[6]_upleveltime75
0.04745	logMelFreqBand_sma[7]_amean
0.03156	logMelFreqBand_sma[7]_kurtosis
0.05312	logMelFreqBand_sma[7]_quartile3
0.01834	logMelFreqBand_sma_de[0]_quartile2
0.03826	logMelFreqBand_sma_de[0]_percentile1.0
0.03894	logMelFreqBand_sma_de[0]_pctlrange0-1
0.02458	logMelFreqBand_sma_de[6]_skewness
0.05487	F0final___Turn_numOnsets
LSP Frequency [0-7] (338→33)	
0.04284	lspFreq_sma[0]_percentile1.0
0.06503	lspFreq_sma[1]_linregerrQ
0.07332	lspFreq_sma[1]_stddev
0.06661	lspFreq_sma[1]_quartile1
0.07334	lspFreq_sma[1]_iqr1-3
0.04470	lspFreq_sma[2]_quartile1
0.08807	lspFreq_sma[3]_skewness
0.06401	lspFreq_sma[4]_linregerrA
0.08617	lspFreq_sma[4]_quartile1
0.06283	lspFreq_sma[4]_quartile2
0.06487	lspFreq_sma[4]_iqr2-3
0.07114	lspFreq_sma[4]_iqr1-3
0.06715	lspFreq_sma[6]_linregc2
0.07958	lspFreq_sma[6]_kurtosis
0.10540	lspFreq_sma[6]_quartile2
0.08661	lspFreq_sma[6]_quartile3
0.09602	lspFreq_sma[6]_iqr1-2
0.08086	lspFreq_sma[7]_linregc2
0.10134	lspFreq_sma[7]_linregerrA
0.07141	lspFreq_sma[7]_kurtosis
0.10714	lspFreq_sma[7]_quartile2
0.09794	lspFreq_sma[7]_quartile3
0.11439	lspFreq_sma[7]_iqr1-3
0.04315	lspFreq_sma_de[0]_kurtosis
0.01422	lspFreq_sma_de[1]_quartile2
0.06444	lspFreq_sma_de[1]_pctlrange0-1
0.00925	lspFreq_sma_de[2]_amean
0.02030	lspFreq_sma_de[2]_kurtosis
0.01095	lspFreq_sma_de[4]_upleveltime75
0.08524	lspFreq_sma_de[7]_linregc1
0.09807	lspFreq_sma_de[7]_linregerrA
0.08175	lspFreq_sma_de[7]_quartile3
0.05487	F0final___Turn_numOnsets
PCM loudness (44→5)	
0.03092	pcm_loudness_sma_stddev
0.04816	pcm_loudness_sma_quartile1
0.06874	pcm_loudness_sma_percentile1.0
0.00284	pcm_loudness_sma_de_upleveltime75
0.05487	F0final___Turn_numOnsets
F0 features (124→17)	
0.24179	F0finEnv_sma_amean
0.2317	F0finEnv_sma_linregerrA
0.22036	F0finEnv_sma_stddev
0.17025	F0finEnv_sma_quartile1
0.28749	F0finEnv_sma_quartile2
0.2569	F0finEnv_sma_quartile3

0.18045	F0finEnv_sma_de_linregc2
0.19299	F0finEnv_sma_de_stddev
0.16265	F0finEnv_sma_de_iqr1-3
0.00899	voicingFinalUnclipped_sma_de_quartile2
0.26651	F0final_sma_amean
0.24975	F0final_sma_linregc2
0.23616	F0final_sma_quartile1
0.28905	F0final_sma_quartile2
0.25133	F0final_sma_quartile3
0.18978	F0final_sma_de_stddev
0.22813	F0final_sma_de_percentile99.0
Jitter and Shimmer (116→20)	
0.1125	jitterLocal_sma_linregerrA
0.0918	jitterLocal_sma_iqr2-3
0.0702	jitterLocal_sma_upleveltime90
0.1272	jitterDDP_sma_amean
0.1256	jitterDDP_sma_linregerrA
0.0681	jitterDDP_sma_quartile1
0.1184	jitterDDP_sma_quartile3
0.1257	jitterDDP_sma_iqr2-3
0.1106	jitterDDP_sma_iqr1-3
0.1191	jitterDDP_sma_percentile99.0
0.0809	jitterDDP_sma_upleveltime75
0.0883	jitterDDP_sma_upleveltime90
0.0699	shimmerLocal_sma_skewness
0.0761	jitterLocal_sma_de_linregc1
0.1078	jitterLocal_sma_de_percentile99.0
0.1267	jitterDDP_sma_de_linregerrA
0.1253	jitterDDP_sma_de_stddev
0.0979	jitterDDP_sma_de_quartile1
0.0842	shimmerLocal_sma_de_linregc1
0.0725	shimmerLocal_sma_de_percentile99.0
All features (1582→80)	
0.02150	mfcc_sma[1]_kurtosis
0.09726	mfcc_sma[3]_percentile1.0
0.04740	mfcc_sma[3]_percentile99.0
0.02303	mfcc_sma[4]_kurtosis
0.06009	mfcc_sma[5]_quartile1
0.08968	mfcc_sma[6]_amean
0.03711	mfcc_sma[7]_quartile2
0.03659	mfcc_sma[8]_linregc2
0.08453	mfcc_sma[9]_quartile3
0.01415	mfcc_sma[9]_upleveltime90
0.08893	mfcc_sma[10]_quartile2
0.02100	mfcc_sma[11]_maxPos
0.04579	mfcc_sma[11]_minPos
0.17968	mfcc_sma[11]_amean
0.15975	mfcc_sma[11]_skewness
0.17976	mfcc_sma[11]_quartile1
0.12959	mfcc_sma[11]_iqr1-2
0.17596	mfcc_sma[11]_percentile1.0
0.13702	mfcc_sma[11]_percentile99.0
0.05280	mfcc_sma[12]_linregc1
0.02626	mfcc_sma[12]_quartile3
0.02522	mfcc_sma[12]_pctlrange0-1
0.13081	mfcc_sma[13]_quartile3
0.20975	mfcc_sma[14]_amean
0.14624	mfcc_sma[14]_linregc2
0.21614	mfcc_sma[14]_quartile1
0.20331	mfcc_sma[14]_quartile2
0.19106	mfcc_sma[14]_quartile3
0.16419	mfcc_sma[14]_iqr1-2
0.20589	mfcc_sma[14]_percentile1.0
0.14288	mfcc_sma[14]_percentile99.0
0.06160	logMelFreqBand_sma[0]_percentile1.0
0.06035	logMelFreqBand_sma[3]_upleveltime75
0.05492	logMelFreqBand_sma[6]_stddev
0.04284	lspFreq_sma[0]_percentile1.0
0.07333	lspFreq_sma[1]_stddev
0.08807	lspFreq_sma[3]_skewness
0.08617	lspFreq_sma[4]_quartile1
0.03253	lspFreq_sma[5]_upleveltime90
0.11472	lspFreq_sma[7]_amean
0.24179	F0finEnv_sma_amean
0.23170	F0finEnv_sma_linregerrA
0.22036	F0finEnv_sma_stddev
0.17025	F0finEnv_sma_quartile1
0.28749	F0finEnv_sma_quartile2
0.25690	F0finEnv_sma_quartile3
0.01363	mfcc_sma_de[1]_quartile2
0.01287	mfcc_sma_de[2]_skewness
0.00525	mfcc_sma_de[3]_quartile2
0.01438	mfcc_sma_de[4]_amean
0.03349	mfcc_sma_de[4]_percentile99.0
0.00543	mfcc_sma_de[5]_skewness
0.00887	mfcc_sma_de[6]_skewness
0.00536	mfcc_sma_de[7]_upleveltime75
0.00314	mfcc_sma_de[8]_upleveltime75
0.00476	mfcc_sma_de[9]_skewness
0.00935	mfcc_sma_de[10]_skewness
0.02548	mfcc_sma_de[11]_skewness
0.01307	mfcc_sma_de[12]_skewness

0.00580	mfcc_sma_de[13]_kurtosis	0.19299	F0finEnv_sma_de_stddev
0.00630	mfcc_sma_de[13]_quartile2	0.00903	voicingFinalUnclipped_sma_de_quartile2
0.00964	mfcc_sma_de[14]_skewness	0.26651	F0final_sma_amean
0.03826	logMelFreqBand_sma_de[0]_percentile1.0	0.24975	F0final_sma_linregc2
0.02458	logMelFreqBand_sma_de[6]_skewness	0.23616	F0final_sma_quartile1
0.04315	lspFreq_sma_de[0]_kurtosis	0.28905	F0final_sma_quartile2
0.02714	lspFreq_sma_de[0]_percentile1.0	0.25133	F0final_sma_quartile3
0.02030	lspFreq_sma_de[2]_kurtosis	0.18978	F0final_sma_de_stddev
0.01095	lspFreq_sma_de[4]_upleveltime75	0.22813	F0final_sma_de_percentile99.0
0.03461	lspFreq_sma_de[6]_kurtosis	0.12854	jitterDDP_sma_de_linregerrA
0.18045	F0finEnv_sma_de_linregc2		

My Research Publications

1. P. Nguyen, T. Le, D. Tran, X. Huang, and D. Sharma. **Fuzzy Support Vector Machines for Age and Gender Classification**. International Conference on Spoken Language Processing (INTERSPEECH) 2010 (ERA rank: A)

2. D. Tran, W. Ma, D. Sharma, and P. Nguyen. **Fuzzy Feature Weighting Techniques for Vector Quantisation**. IEEE World Congress on Computational Intelligence (WCCI) 2010 (ERA rank: A)

3. T. Hoang, P. Nguyen, T. Le, D. Tran, and D. Sharma. **Enhancing Performance of SVM-Based Brain-Computer Interface Systems**. 17th International Conference on Neural Information Processing (ICONIP) 2010, Australian Journal of Intelligent Information Processing Systems (AJIIPS Journal) 2010 (ERA rank: A)

4. P. Nguyen, D. Tran, X. Huang, and D. Sharma. **Automatic Speech-based Classification of Gender, Age and Accent**. The 11th International Workshop on Knowledge Management and Acquisition for Smart Systems and Services (PKAW) 2010 (ERA rank: B)

5. P. Nguyen, D. Tran, X. Huang, and D. Sharma. **Australian Accent-Based Speaker Classification**. The 3rd International Conference on Knowledge Discovery and Data Mining (WKDD) 2010 (ERA rank: C)

6. P. Nguyen, D. Tran, X. Huang, and D. Sharma. **Automatic Classification of Speaker Characteristics**. The 3rd International Conference on Communications and Electronics (HUT ICCE) 2010

Bibliography

[1] K. Amilon, J. Weijer, and S. Schötz. **The Impact of Visual and Auditory Cues in Age Estimation**. In Christian Müller, editor, *Speaker Classification II*, **4441** of *Lecture Notes in Computer Science*, pages 10–21. Springer Berlin / Heidelberg, 2007.

[2] K. Ananthakrishnan, A. Hashemi-Sakhtsari, A. Barnes, S. Bailes, C. Watson, and P. Warren. **Performance of speaker-independent speech recognisers for automatic recognition of Australian english**, 2006.

[3] Language Australia and Identity in. **Language and Identity in Australia**, 2000.

[4] A. Batliner and R. Huber. **Speaker Characteristics and Emotion Classification**. In Christian Müller, editor, *Speaker Classification I*, **4343** of *Lecture Notes in Computer Science*, pages 138–151. Springer Berlin / Heidelberg, 2007.

[5] K. Berkling, M. Zissman, J. Vonwiller, and C. Cleirigh. **Improving accent identification through knowledge of English syllable structure**. In *ICSLP-1998*, **2**, pages 89–92, 1998.

[6] T. Bocklet, A. Maier, and E. Nöth. **Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression**. In Petr Sojka, Aleš Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue*, **5246** of *Lecture Notes in Computer Science*, pages 253–260. Springer Berlin / Heidelberg, 2008.

[7] R. Brett and K. Kuldip. **GMM Based Speaker Recognition on Readily Available Databases**, 2003. CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States) ER.

[8] C. Burges. **A tutorial on support vector machines for pattern recognition**. *Data Mining and Knowledge Discovery*, **2**(2):121–167, 1998.

[9] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. **A Database of Age and Gender Annotated Telephone Speech**. In Nicoletta Calzolari Chair, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010.

[10] W. Chao and S. Seneff. **Robust pitch tracking for prosodic modeling in telephone speech**. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, **3**, pages 1343–1346 vol.3, 2000.

[11] T. Chen, C. Huang, E. Chang, and J. Wang. **Automatic Accent Identification Using Gaussian Mixture Models**. In *in IEEE Workshop on ASRU*, pages 343–346, 2001.

[12] C. Chang and C. Lin. **LIBSVM: a Library for Support Vector Machines**, 2001.

[13] G. Choueiter, G. Zweig, and P. Nguyen. **An empirical study of automatic accent classification**, 2008.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. **Emotion recognition in human-computer interaction**. *IEEE Signal Processing Magazine*, **18**(1):32–80, 2002.

[15] J. Deller, J. Hansen, and J. Proakis. *Discete-Time Processing of Speech Signals*. Wiley, N.Y, 2000.

[16] V. Dellwo, M. Huckvale, and M. Ashby. **How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification**. In Christian Müller, editor, *Speaker Classification I*, **4343** of *Lecture Notes in Computer Science*, pages 1–20. Springer Berlin / Heidelberg, 2007.

[17] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. **Emotional speech: Towards a new generation of databases**. *Speech Communication*, **40**(1-2):33–60, 2003.

[18] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[19] E. Eriksson, R. Rodman, and R. Hubal. **Emotions in Speech: Juristic Implications**. In Christian Müller, editor, *Speaker Classification I*, **4343** of *Lecture Notes in Computer Science*, pages 152–173. Springer Berlin / Heidelberg, 2007.

[20] F. Eyben, M. Wöllmer, and B. Schuller. *openSMILE the Munich open Speech and Music Interpretation by Large Space Extraction toolkit*. 2010.

[21] F. Eyben, M. Wöllmer, and B. Schuller. **OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit**. 2009. Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009.

[22] S. Furui. **Recent advances in speaker recognition**. *Pattern Recognition Letters*, **18**(9):859–872, 1997.

[23] R. Gajsek, J. Zibert, T. Justin, V. Struc, B. Vesnicer, and F. Mihelic. **Gender and Affect Recognition Based on GMM and GMM–UBM modeling with relevance MAP estimation**. In Isca, editor, *Proceedings of Interspeech*, 2010.

[24] G. Garcia, S. Jung, and T. Eriksson. **Bayes-optimal estimation of GMM parameters for speaker recognition**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4441 LNAI**:142–156, 2007.

[25] U. Gut. **Foreign accent**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4343 LNAI**:75–87, 2007.

[26] M. Hall. **Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning**, 2000.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and Ian Witten. **The WEKA data mining software: an update**. *SIGKDD Explor. Newsl.*, **11**(1):10–18, 2009.

[28] J. Harrington, F. Cox, and Z. Evans. **An acoustic phonetic study of broad, general, and cultivated Australian English vowels**. *Australian Journal of Linguistics*, **17**(2):155 – 184, 1997.

[29] R. Hathaway. **Another interpretation of the EM algorithm for mixture distributions**. *Statistics and Probability Letters*, **4**(2):53–56, 1986.

[30] D. Hill. **Speaker Classification Concepts: Past, Present and Future**. In Christian Müller, editor, *Speaker Classification I*, **4343** of *Lecture Notes in Computer Science*, pages 21–46. Springer Berlin / Heidelberg, 2007.

[31] X. Huang, K. Lee, H. Hon, and M. Hwang. **Improved acoustic modeling with the SPHINX speech recognition system**. **1**, pages 345–348, 1991. Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.

[32] C. Ishi, H. Ishiguro, and N. Hagita. **Using Prosodic andVoice Quality Features for Paralinguistic InformationExtraction**. In *Speech Prosody 2006*, 2006.

[33] M. Kockmann, L. Burget, and J. Černocký. **Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge**. In Isca, editor, *Proceedings of Interspeech*, 2010.

[34] K. Kumpf and R. W. King. **Automatic accent classification of foreign accented Australian English speech**. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, **3**, pages 1740–1743 vol.3, 1996.

[35] C. Lee and S. Narayanan. **Toward detecting emotions in spoken dialogs**. *Speech and Audio Processing, IEEE Transactions on*, **13**(2):293–303, 2005.

[36] H. Meinedo and I. Trancoso. **Age and Gender Classification using Fusion of Acoustic and Prosodic Features**. In Isca, editor, *Proceedings of Interspeech*, 2010.

[37] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. Bauer, and B. Littel. **Comparison of four approaches to age and gender recognition for telephone applications**. **4**, pages IV1089–IV1092, 2007. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.

[38] J. Millar, J. Vonwiller, J. Harrington, and P. Dermody. **The Australian National Database of Spoken Language**. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, **i**, pages I/97–I100 vol.1, 1994.

[39] N. Minematsu, M. Sekiguchi, and K. Hirose. **Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers**. **1**, pages I/137–I/140, 2002. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.

[40] A. Mitchell. *The pronunciation of English in Australia / by A.G.Mitchell and A.Delbridge*. A. & R, Syd. :, 1965. (Alexander George) Rev. (i.e. 2nd) ed.

[41] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler. **Floating search methods for feature selection with nonmonotonic criterion functions**. In *Pattern Recognition,*

1994. Vol. 2 - Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on, **2**, pages 279–283 vol.2, 1994.

[42] D. Reynolds. **An overview of automatic speaker recognition technology**. **4**, pages IV/4072–IV/4075, 2002. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.

[43] D. Reynolds. **Speaker identification and verification using Gaussian mixture speaker models**. *Speech Commun.*, **17**(1-2):91–108, 1995.

[44] D. Reynolds, T. Quatieri, and R. Dunn. **Speaker Verification Using Adapted Gaussian Mixture Models**. *Digital Signal Processing*, **10**(1-3):19–41, 2000.

[45] D. Reynolds and R. Rose. **Robust text-independent speaker identification using Gaussian mixture speaker models**. *IEEE Transactions on Speech and Audio Processing*, **3**(1):72–83, 1995.

[46] S. Schacht, J. Koreman, C. Lauer, A. Morris, D. Wu, and D. Klakow. **Frame Based Features**. In Christian Müller, editor, *Speaker Classification I*, **4343** of *Lecture Notes in Computer Science*, pages 226–240. Springer Berlin / Heidelberg, 2007.

[47] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. **The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals**. **2**, pages 881–884, 2007. International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007.

[48] B. Schuller, R. Müller, M. Lang, and G. Rigoll. **Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles**. In *Interspeech 2005 - Eurospeech*, Lisbon, Portugal, 2005.

[49] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. **Speaker independent speech emotion recognition by ensemble classification**. **2005**, pages 864–867, 2005.

[50] B. Schuller, S. Reiter, and G. Rigoll. **Evolutionary feature generation in speech emotion recognition**. **2006**, pages 5–8, 2006.

[51] B. Schuller, S. Steidl, and A. Batliner. **The INTERSPEECH 2009 emotion challenge**. pages 312–315, 2009. Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009.

[52] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. **The INTERSPEECH 2010 Paralinguistic Challenge**. In Isca, editor, *Proceedings of Interspeech*, 2010.

[53] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. **Acoustic emotion recognition: A benchmark comparison of performances**. In *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009.*, pages 552–557, 2009. Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009. No.: 5372886.

[54] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll. **Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?** pages 4501–4504, 2008.

[55] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll. **Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs**. In Sylvie Hancil, editor, *The Role of Prosody in Affective Speech*, **97** of *Linguistic Insights, Studies in Language and Communication*. Peter Lang Publishing Group, 2009.

[56] T. Schultz. **Speaker characteristics**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4343 LNAI**:47–74, 2007.

[57] S. Schötz. **Acoustic analysis of adult speaker age**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4343 LNAI**:88–107, 2007.

[58] S. Schötz and C. Müller. **A Study of Acoustic Correlates of Speaker Age**. In Christian Müller, editor, *Speaker Classification II*, **4441** of *Lecture Notes in Computer Science*, pages 1–9. Springer Berlin / Heidelberg, 2007.

[59] I. Shafran, M. Riley, and M. Mohri. **Voice signatures**, 2003.

[60] M. Shami and W. Verhelst. **Automatic classification of expressiveness in speech: A multi-corpus study**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4441 LNAI**:43–56, 2007.

[61] E. Shriberg. **Higher-level features in speaker recognition**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4343 LNAI**:241–259, 2007.

[62] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. PhD thesis, 2009.

[63] D. Sturim, W. Campbell, and D. Reynolds. **Classification methods for speaker recognition**, 2007.

[64] D. Tran. *Fuzzy Approaches to Speech and Speaker Recognition*. PhD thesis, 2000.

[65] D. Tran, W. Ma, D. Sharma, and T. Nguyen. **Fuzzy vector quantization for network intrusion detection**. pages 566–570, 2007. Proceedings - 2007 IEEE International Conference on Granular Computing, GrC 2007.

[66] D. Ververidis and C. Kotropoulos. **Emotional speech recognition: Resources, features, and methods**. *Speech Communication*, **48**(9):1162–1181, 2006.

[67] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. **On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition**. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, Roberto Pieraccini, and Michael Weber, editors, *Perception in Multimodal Dialogue Systems*, **5078** of *Lecture Notes in Computer Science*, pages 217–220. Springer Berlin / Heidelberg, 2008.

[68] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. **Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing**. In Ana Paiva, Rui Prada, and Rosalind Picard, editors, *Affective Computing and Intelligent Interaction*, **4738** of *Lecture Notes in Computer Science*, pages 139–147. Springer Berlin / Heidelberg, 2007.

[69] P. Woodland, M. Gales, D. Pye, and S. Young. **Broadcast news transcription using HTK**. **2**, pages 719–722, 1997.

[70] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.4)*. 2009.

[71] G. Zweig, Y. Ju, P. Nguyen, D. Yu, Y. Wang, and A. Acero. **Voice-rate: a dialog system for consumer ratings**, 2007.