Automatic Gender Recognition Based on Audio-Visual Cues

MARIANNA MADRY-PRONOBIS

KTH Electrical Engineering

Master's Degree Project
Stockholm, Sweden May 2009

XR-EE-SIP 2009:006

Abstract

The ability to perform automatic recognition of human gender is important for a number of systems that process or exploit human-source information. The outcome of an Automatic Gender Recognition (AGR) system can be used for improving intelligibility of man-machine interactions, annotating video files or reducing the search space in subject recognition or surveillance systems. In the previous studies, the AGR systems were typically based on only one modality (audio or vision) and their robustness in real-world scenarios was seldom considered. However, in many typical applications, both audio signal and visual signal are available. Ideally, an AGR system should be able to exploit both modalities to improve the overall robustness. In this work, we develop a multi-modal AGR system based on audio and visual cues and present its thorough evaluation in realistic scenarios. First, in the framework of two uni-modal AGR systems, we analyze robustness of different audio features (pitch frequency, formant and cepstral representations) and visual features (eigenfaces, fisherfaces) under varying conditions. Then, we build an integrated audio-visual system by fusing information from each modality at the classifier level. Additionally, we evaluate performance of the system with respect to quality of data used for training the system. We conducted the AGR studies on the BANCA database. In the framework of the uni-modal AGR systems, we show that: (a) the audio-based system is more robust than the vision-based system and its resilience to noisy conditions is increased by modelling only voiced speech frames; (b) in case of audio, the cepstral features are superior to the pitch frequency and formant features, and in case of vision, the fisherfaces outperforms the eigenfaces; (c) for the cepstral features, modelling of higher spectral details and the use of both static and delta coefficients makes the system robust towards noisy conditions. The integration of audio and visual cues yields a robust system that preserves the performance of the best modality in clean conditions and helps in improving performance in noisy conditions. Finally, the multi-conditional training (clean+noisy data) helps in improving performance of the visual features and, consequently, the recognition rate of the audio-visual AGR system.

Acknowledgements

This MSc project has been made possible with the financial support of the EU 6th framework IST *Augmented Multi-party Interaction with Distance Access (AMIDA)* project that focuses on developing technologies for supporting meetings from a distance, such as video or phone conferences.

First and foremost, I would like to express my sincere gratitude to my supervisors, Mathew Magimai.-Doss at the Idiap Research Institute and Professor Arne Leijon at the Royal Institute of Technology (KTH). I own many thanks to Mathew for leading me into the world of speech processing and recognition, and teaching step-by-step how to do research. Thank you for your guidance, great enthusiasm, support and friendship. I am also very grateful to Arne for his valuable advices and suggestions, support and kindness. The visual part of the project has been made possible with the help and suggestions received from Sébastien Marcel and Luo Jie. Finally, I wish to acknowledge Vladislav Nenchev who acted as the opponent on this thesis and provided valuable comments regarding the report.

Spacial thanks to all my friends at Idiap for an absolutely fantastic time spend together, for skiing, trekking, traveling and much more. My stay in Martigny would not be the same without you: Barbara, Bogdan, Chris, Constantin, Danil, Elisa, Ferran, Francesco, Ganga, Ghita, Giulia, Guillermo, Hamed, Joseph, Kate, Luo Jie, Niklas, Petr, Radu, Stefan, Tatiana, Tristan and Vincent. I would also like to thank all my fellow friends at KTH for their enthusiasm and unforgettable atmosphere: Adam, Alessandro, Chithrupa, Davide, Matteo and Vlad. Last but certainly not least, I want to thank Andrzej for always being with me and for me.

Marianna Pronobis
Stockholm, May 2009

.

Contents

Contents **v**

List of Figures **vii**

List of Tables **ix**

1 Introduction **1**
1.1 Related Work 1
1.2 Main Objectives 2
1.3 Contribution of the Thesis 3
1.4 Outline 4

2 Automatic Gender Recognition **5**
2.1 Problem Statement 5
2.1.1 Uni-modal AGR System 7
2.2 Audio Features 8
2.2.1 Voice Source Related Features 11
2.2.2 Vocal Tract Related Features 11
2.3 Visual Features 18
2.3.1 Eigenfaces 19
2.3.2 Fisherfaces 21
2.4 Classification 22
2.4.1 Support Vector Machine 23
2.5 Cue Integration 26
2.5.1 Audio-Visual AGR System 30

3 Audio-Based AGR Studies **33**
3.1 Motivation and Objectives 33
3.1.1 Experiments 34
3.2 Experimental Setup 34
3.2.1 Database 35
3.2.2 Analysis of Audio Data 35
3.2.3 Classification 36
3.2.4 Performance Evaluation and Cue Integration 36

3.3 Results and Discussion 37
3.3.1 Frame Selection 37
3.3.2 Voice Source Related Features 38
3.3.3 Vocal Tract Related Features 40
3.3.4 Audio Cue Integration 45
3.4 Summary and Conclusions 47

4 Vision-Based AGR Studies **49**
4.1 Motivation and Objectives 49
4.2 Experimental Setup 50
4.2.1 Database 50
4.2.2 Analysis of Visual Data 51
4.2.3 Classification 52
4.2.4 Performance Evaluation 52
4.3 Results and Discussion 53
4.4 Summary and Conclusions 54

5 Audio-Visual AGR Studies **55**
5.1 Motivation and Objectives 55
5.2 Comparision of Audio and Visual Features 56
5.3 Experimental Setup 57
5.3.1 Database 57
5.3.2 System Setup 58
5.3.3 Performance Evaluation 58
5.4 Results and Discussion 59
5.5 Summary and Conclusions 61

6 Conclusions **63**
6.1 Future Work 64

A BANCA Database **65**

Bibliography **69**

List of Figures

2.1 Overview of the architecture of the uni-modal AGR system 7
2.2 Model of speech production process (from [45]). 9
2.3 Relation between the true frequency in *Hz* and the frequency perceived by humans according to the mel scale. 16
2.4 Triangular filters for transforming the frequency axis of the short-term speech spectrum onto the mel-scale (from [14]). 16
2.5 Overview of the architecture of the AV-AGR system 31

3.1 Distributions of fundamental frequency for females and males 39
3.2 Performance of the A-AGR system for the cepstral features across protocols **0** and **A** 42
3.3 Performance of the A-AGR system for the cepstral features across protocols **A** and **B** 43
3.4 Performance of the A-AGR system for the cepstral features across protocols **A** and **C** 44
3.5 Performance of the A-AGR system through the audio cue integration experiments 46

4.1 Examples of images from the BANCA database 51
4.2 Examples of automatically detected face regions in the BANCA images 51

5.1 Performance of the AV-AGR system with respect to the integration rule 59
5.2 Performance of the AV-AGR system with respect to the type of weighting 60
5.3 Performance of the uni-modal AGR systems and the AV-AGR system . 60

A.1 Examples of images from the BANCA database 66

List of Tables

3.1 Experimental setup for protocols **0**, **A**, **B** and **C** for an audio part of the BANCA database. 36
3.2 Performance of the A-AGR system through the frame selection experiments 38
3.3 Performance of the A-AGR system for the voice source related features 39
3.4 Performance of the A-AGR system for the voice source and vocal tract related features 40
3.5 Performance of the A-AGR system through the audio cue integration experiments 46
3.6 Importance of fundamental frequency in the correct classification when integrated with the cepstral features 46

4.1 Experimental setup for protocols **0**, **A**, **B** and **C** for a visual part of the BANCA database 52
4.2 Performance of the V-AGR system for the eigenfaces and fisherfaces . . 53

5.1 Performance of the A-AGR and V-AGR system for selected features. . . 57
5.2 Performance of the AV-AGR system 62
5.3 Importance of the audio features in the correct classification when integrated with the visual features 62

A.1 Division of subjects in the BANCA database 66
A.2 Experimental setup for protocols **0**, **A**, **B** and **C** for the BANCA database. 67

Chapter 1

Introduction

The ability to perform automatic recognition of human gender is crucial for a number of systems that process or exploit human-source information. Typical examples are information retrieval, human-computer or human-robot interaction. The outcome of an *Automatic Gender Recognition* (AGR) system can be used for generating meta-data information useful for annotating audio and video files. Moreover, gender is an important cue that can be exploited for improving intelligibility of man-machine interaction, or simply, for reducing the search space in applications such as speaker recognition or surveillance systems.

The problem of AGR has been addressed in the past by several authors (see Section 1.1) using only one modality (audio or vision). The investigations were performed mainly under clean conditions and the robustness of AGR systems in real-world scenarios was seldom considered. However, in many typical applications, both audio and vision are available. Ideally, an AGR system should be able to exploit both modalities to improve robustness. Since each modality has different characteristics, audio-visual cues can provide a more comprehensive description of a subject than a single modality. Finally, integration of the cues may yield a AGR system that is resilient to the degradation of both, or even to temporal unavailability of one of the input signals.

The goal of this work is to develop a multimodal AGR system based on audio and visual cues that is to be robust under varying conditions that occur in realistic scenarios.

1.1 Related Work

In this section, the previous work on automatic gender recognition is briefly reviewed. More information about methods and results obtained in the previous studies can be found in Chapter 2.

The previously proposed solutions to the AGR problem were based on single modality, i.e. either on audio or vision. The first works on audio-based AGR mainly aimed at identifying appropriate features of speech signal for the task. In

particular, comparison of voice source- (pitch frequency) and vocal tract-related features (first four formants with their respective frequency, amplitude and bandwidth) for ten vowels extracted from the clean-condition speech data was presented in [10]. Further analysis of different parametric representations of speech signal (linear prediction, autocorrelation, reflection and cepstral) was performed on the same database for vowels, voiced and unvoiced fricatives [69]. The evaluation of mel-cepstral features for different groups of phonemes like vowels, nasal, liquids etc. was conducted in [18]. More recently, the comparision of different types of classifiers (such as nearest neighbor classifier, Support Vector Machine (SVM)) for the cepstral coefficients was presented on high quality recordings from the ISOLET corpus [68].

Early research in vision-based gender recognition was focussed upon the use of artificial neural networks for feature extraction and classification on clean condition data [13, 20]. Subsequently, the applicability of geometrical features (such as eyebrow thickness or nose width) indicated by psychological studies on gender recognition by humans was verified [7]. Latest research looked into more complex lighting and pose variations, and for larger sets of subjects, such as in the FERET database [39]. The performance of different types of classifiers for the AGR task based on visual cues was studied, such as the linear, quadratic, fisher linear discriminant, k-nearest neighbor classifiers as well as the more complex techniques such as SVMs or large ensemble Radial Basis Function (RBF) networks [39, 68]. In [68], comparison of row data reprasentation with features obtained through principal component analysis (PCA), referred to as eigenfaces [56], was made on database consisting of frontal, un-occluded face images.

Motivations for this work can be found, inter alia, in limitation of the earlier studies. First, the aforementioned researches utilized only one modality (audio or vision). The comparison of the performance of audio and visual features presented in [68] was done on different databases thus limiting the interpretation of the obtained results. To the best knowledge of the authors, no solutions to the AGR problem based on integrated audio-visual cues were published. Second, the investigations were performed mainly under clean conditions and the robustness of AGR systems in real-world scenarios was seldom considered for both audio and visual cues.

1.2 Main Objectives

The main goal of this work is to develop an audio-visual AGR system that can provide sufficient robustness under varying conditions that occur in realistic scenarios. Typically, during studies on a multimodal AGR a number of issues can be addressed, such as a choice of robust representation of the audio and visual signal (feature selection), accurate classification method, efficient and effective cue integration method or strategy of training. However, motivated by the previous studies on the topic, in this thesis we focus on aspects related with selection of the audio and visual signal representation and the audio-visual cue integration method.

In particular, we would like to identify which audio and visual features will yield a better AGR system and how much performance of the AGR system can be improved by combining these two modalities in a realistic scenario. During the audio-visual AGR studies, we address the following practical questions:

1. *What is the effect of varying conditions on performance of different audio and visual features?* Which audio and visual features are the most robust in realistic scenarios? Which parameters of particular features are crucial for their robustness?

2. *How does the choice of an audio-visual integration method influence performance of the AGR system?* What is the most effective and efficient method of integrating audio and visual cues?

3. *What is the effect of integrating audio and visual information on the AGR system accuracy?* How much does the audio-visual integration help improving performance of the AGR system under varying conditions? How much are the selected audio and visual features complementary? Which type of cues (audio or visual) is more important in correct classification?

4. *What is the effect of training data conditions on the AGR system accuracy?* Is a better strategy to train the AGR system on clean-condition or multi-condition (clean+noisy) data?

The further objectives for the audio, visual and integrated audio-visual AGR studies are stated in Sections 3.1, 4.1, and 5.1, respectively.

1.3 Contribution of the Thesis

In this thesis, we investigate an multimodal AGR system based on audio and visual cues. First, in the framework of two uni-modal AGR systems, we analyze robustness of different audio (pitch frequency, formant and cepstral representations) and visual (eigenfaces, fisherfaces) features under varying conditions. Then, we build an integrated audio-visual system by fusing information from each modality at the classifier level and we study various cue integration methods. Additionally, we evaluate performance of the system with respect to quality of data used for system training. Our studies were conducted on the BANCA database [16] comprising datasets of varying complexity, and in the basic setup, the AGR system was trained exclusively on clean and tested on clean or noisy data. In the framework of the uni-modal AGR systems, we show that:

- the audio-based system is more robust than the vision-based system and its resilience to noisy conditions is increased by modelling only voiced speech frames;

- in case of audio, the cepstral features are superior to the pitch frequency and formant features, and in case of vision, the fisherfaces outperforms the eigenfaces;
- for the cepstral features, modelling of higher spectral details and the use of both static and delta coefficients makes the system robust towards noisy conditions.

The integration of audio and visual cues yields a robust system that preserves the performance of the best modality in clean conditions and helps in improving performance in noisy conditions. Finally, the multi-conditional training (clean+noisy data) highly improves performance of the visual features and, in consequence, the recognition rate of the audio-visual AGR system. The selection of the audio and visual features and integration of the multi-modal cues result in the resilient AGR system applicable in practical applications.

1.4 Outline

The thesis is organized as follows. Chapter 2 formulates the task of automatic gender recognition as a pattern recognition problem, presents consecutive parts of the uni-modal AGR systems with focus on selection of audio and visual features, and finally, discusses different cue integration methods and presents architecture of the audio-visual AGR system. The experimental studies on the AGR problem for audio, visual and integrated audio-visual cues are presented and discussed in in Chapter 3, 4 and 5, respectively. These chapters have similar structure. First, motivation of the studies is discussed and the main objectives are defined. Then, experimental setup is specified and, finally, the results obtained under varying conditions are presented and discussed. The thesis is summarized with conclusions and suggestions for further work in Chapter 6 and supplemented with the description of the BANCA database in Appendix A.

Publication Parts of the work presented in this thesis has appeared in the following publications:

- M. Pronobis and M. Magimai.-Doss. Integrating audio and vision for robust automatic gender recognition. IDIAP Technical Report, Idiap-RR-73-2008, Idiap, November 2008 [50].
- M. Pronobis and M. Magimai.-Doss. Analysis of F0 and cepstral features for robust automatic gender recognition. IDIAP Technical Report, 2009 [51].

Chapter 2

Automatic Gender Recognition

This chapter discusses the problem of *Automatic Gender Recognition (AGR)* and presents an architecture of an audio-visual AGR system. Section 2.1 formulates the task of automatic gender recognition as a pattern classification problem. Further, the motivation for choosing audio and visual modalities as a source of information about a subject is provided. In designing the audio-visual AGR system a two-step approach is adopted. First, the two modalities are studied separately by building uni-modal AGR systems, namely the audio-based and the vision-based AGR systems. The overview of an architecture of the uni-modal AGR systems is presented in Section 2.1.1. The feature selection from the point of robustness under varying conditions is discussed for both audio and visual signal in Sections 2.2 and 2.3, respectively. In Section 2.4, different types of classification methods are discussed and a description of the classifier used in the system, the *Support Vector Machine (SVM)*, is provided. Finally, the two uni-modal systems are integrated to provide the final decision based on both modalities. Section 2.5 discusses details of audio-visual AGR system based on classifier fusion approach.

2.1 Problem Statement

Automatic Gender Recognition (AGR) can be described as a process of identifying subject's sex given biometric information about a person. The idea is to assign an unambiguous label: female or male to a subject based on information about person's individual attributes. The whole action should involve as little human interaction as possible.

Although there exist large variations in physical appearance and behavior among people of the same gender, all of them possess some characteristic features that indicate their sex. Due to this fact, information about particular person, after being captured by our senses and processed in a brain, may be used to assign an unknown person to the one of gender categories, i.e. may be used to identify the person as a female or male. The correct decision can not be made without some prior knowledge about the typical features for each gender which constitute archetypes

(typical models) of a female and of a male in our consciousness.

The problem presented above is a typical example of a *pattern classification* task where a given object is assigned to one of the predefined categories (also referred to as *classes*) given an observation. Each of the classes includes a group of items of similar properties characterized by features and is represented by a *model* learned from these features. A model can be seen as a description of the features that are common within a class and different between classes. Another problem is to define a proper structure of the model and to automatically train it, i.e. estimating the values of the model parameters for each of the classes. The most important properties of each class should be rendered and a constraint that all models have to be distinguishable from each other should be fulfilled. This task can be solved by introducing the so called *training* phase. A pre-collected set of samples also referred to as *training* data is used to approximate the best parameters of the models. If each sample in the training data was previously assigned to the classes, the problem of *supervised learning* is considered. Otherwise, the class for each sample has to be determined during training. Such approach is referred as *unsupervised learning* or *clustering*. The system should be able not only to classify correctly the training examples, but also new examples that will be introduced to the system in the future. The ability to classify the novel patterns that were not available in training data is known as *generalization*. Thus, in order to obtain models with generalization properties an additional set of data (*development* set) that were not used during training, i.e. in determining parameters of model, is employed. A number of requirements have to be satisfied to ensure reliability of the estimated prototypes. More detailed information regarding this topic can be found in [5, p. 2] and [22, p. 65]. Given the trained model and unseen (test) sample, the *recognition* process consists of feature extraction followed by matching against the model of each class. The final decision is then based on the best matched class.

A gender of a subject belongs to *external* characteristics of an individual (similarly to age or race) and is encoded in both physiological (such as fingerprint, iris scan, DNA code, face or body image, voice pitch etc.) and behavioral (such as signature, gait, typing rhythm, voice timbre and tone etc.) biometrics [28]. In context of AGR problem, all biometrics can be graded with respect to: (a) a degree of subject's attention and cooperation which is needed to collect them, and (b) a degree of differentiation between two genders which they provide. For example, a gender of a subject can be rather poorly judged based on keystrok recordings, but almost perfectly on a DNA sample and very well on subject's voice recordings. At the same time, some of the biometrical signals are relatively easy to collect like e.g. audio or video recordings, where others required subject's cooperation when collected, such as fingerprints or DNA samples[1]. In this work, the AGR system uses audio and visual signal that are non-intrusive and can well discriminate with respect to gender.

[1]When subjects provide data voluntary.

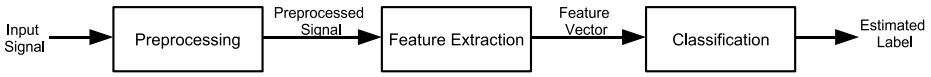


Figure 2.1. Overview of the architecture of the uni-modal AGR system.

2.1.1 Uni-modal AGR System

As mentioned in the previous section, the AGR system is based on two modalities: audio and vision. In our approach, the two modalities are first studied separately by building uni-modal AGR system, namely audio-based AGR system (A-AGR) and vision-based AGR (V-AGR) system. In this section the overview of the uni-modal systems is given.

First, audio and visual information are extracted from a video file. The input to the V-AGR is constituted by a series of images collected from a video which capture a view of a subject. Images are processed sequentially by the system, one by one. Similarly, the A-AGR works based on an audio signal containing recording of a subject's voice. The audio stream is divided into short frames ($10 - 20ms$) in order to ensure stationary properties of the speech signal. The consecutive frames are processed sequentially. In the proposed solution, two main assumptions are made. First, the functionality of the AGR system is limited to gender recognition of only a single subject at a particular time instance. This requirement excludes from further consideration such phenomena and situations like cross-talking in the audio system, or simply, occurrence of two or more people in an image. Second, the V-AGR system exploits exclusively face images and no stature information is used.

Both A-AGR and V-AGR systems have similar architecture presented in Figure 2.1. Each of the systems consists of three parts performing the following functions: (a) signal preprocessing, (b) feature extraction, and (c) classification. The role of the signal preprocessing block is extraction of useful fragments of the signal. Previous AGR studies using audio suggest that voiced phonemes are more discriminative for gender than unvoiced phonemes [69, 18]. We use a speech/non-speech or voiced/unvoiced speech detection to obtain the most informative parts of the signal (the method used for voiced/unvoiced detection is described in Section 2.2.1). In case of the V-AGR system, data preprocessing includes face detection, localization, and finally segmentation.

The function of the second block is extraction of features from the preprocessed signal that are good representation for gender classes. The feature extraction for the A-AGR and V-AGR system is discussed in Sections 2.2 and 2.3, respectively. Finally, classification of an instance to one of the two possible classes (female or male) is performed. The classification module employs the same algorithm in case of both A-AGR and V-AGR. The detailed description of classification method is given in Section 2.4.

2.2 Audio Features

In this section, we describe different acoustic features which automatically extracted from the speech signal offer good discrimination between genders, and at the same time ensure robustness of an A-AGR system under varying conditions that occur in realistic scenarios. We start with a short overview of the speech production model, since it is essential to the discussion provided further in the section. Next, we indicate a number of physiological factors that differentiate the female and male voices, and specify their influence on the acoustic parameters of speech signal. Then, the audio signal preprocessing operations are described. In Sections 2.2.1 and 2.2.2 different types of the vocal tract and the voice source related features are presented and their application to AGR problem under varying conditions is discussed.

The vocal system instantiates sounds as air-pressure waves. Air-stream pushed out from the lungs passes through the vocal folds which may work in one of following modes: (a) vibrating i.e. the folds open and close rhythmically causing oscillations of the air-stream; in this way voiced sounds are generated; (b) constantly open, when unvoiced sounds are generated. Later, the air-pressure wave goes through trachea and larynx to the oral, nasal or both cavities. The final shape of the sound wave is formed by different positions and movements of articulators (tongue, teeth, lips etc). The process of speech production, as shown in Figure 2.2, is modeled by the system consisting of two components: *voice (excitation) source* which imitates the influence of vocal cords, and *time-varying filter* which represents the effect of the propagation of the sound through the vocal tract [67, p. 12]. The excitation signal is generated by an *impulse generator*, a *random number generator* or a combination of both, depending on whether the produced sound is voiced, unvoiced or mixed voiced/unvoiced. The impulse generator is characterized by the fundamental frequency of the vocal cords *F0* (also called *pitch frequency*[2]). As mentioned before, the sound wave is further formed during propagation through the trachea, larynx and cavities. The structure of the vocal tract can be described by theory created for the acoustic tube, where during propagation the air pressure waves resonate at different frequencies. The resonating frequencies are are called *formants* (*F1-F4*). The human vocal tract shapes spectral characteristic of the excitation signal according to the frequency response of the acoustic tube. Therefore, the digital filter may be used to represent the influence of the passage of the sound through the vocal system. Resonance frequencies of the acoustic tube (vocal tract) depend on the length and shape of the tube as well as on whether it has open or closed ends. Therefore, due to different configuration of the articulators during production different sound (air pressure waves with different characteristics) are generated. However, even the two acoustic waves of the same sound produced by the same person can not be identical. The 'intra-speaker' variations may be caused by the different physical or mental states of the speaker. Considerably larger variation exists between acous-

[2]Pitch represents the perceived fundamental frequency of a sound.

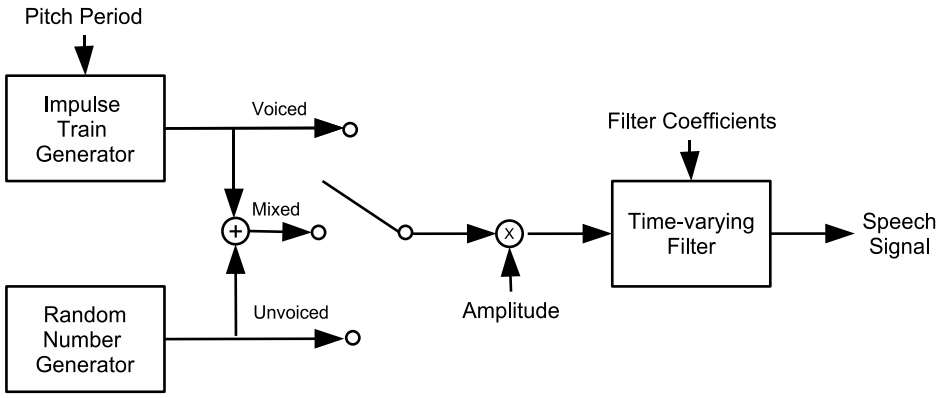


Figure 2.2. Model of speech production process presented by Oppenheim and Schafer in [45, p. 512].

tic waves among different speakers for the same sound. In this case, the variation mostly originate from the anatomy of the speaker's vocal tract, but also depend on other factors, such as speech style and speaking rate [27, p. 4].

Numerous studies have been conducted in order to indicate the factors that differentiate the female and male voice, and in consequence, allow humans for very precise and fast recognition of the speaker gender [12, 37, 52, 59]. Generally, the factors are divided into the two main groups [69]:

1. *objective* factors, that can be directly measured, like physiological differences in the structure of vocal organs;

2. *subjective* factors, that can be only psychophysically assessed, like perceptual features of the voice.

A number of differences in anatomy of female and male vocal tracts were indicated. The most important factor that differs the vocal tracts is their length. The ratio between the total length of the female vocal tract to that of male is around $0.8 - 0.87$ [17, 25]. Further, it was shown that the female larynx differs from the male in vocal fold length, thickness, angle of the thyroid laminae, resting angle of the glottis and vertical convergence angle in the glottis [62]. Regarding the perceptual features of the voices, it has been found that the female voice is typically more breathy and melodic than the male voice, and finally that females typically tend to speak faster than males [30].

The above mentioned physiological and perceptual factors lead to differences in acoustic feature parameters [17, 25, 30, 62, 69]. The factors relate to both the *voice source-* and the *vocal tract-related* parameters, i.e. the dissimilarities are visable both in the pitch period, voicing and amplitude of speech, as well as in the shape of the short-term speech signal spectrum. More specifically, the female and male voice differ in [10, 40, 62, 63, 69]:

- typical value of the pitch frequency – the female pitch frequency is higher than male; typical values for females range between 170-280Hz and for males between 110-150Hz;
- location of formants for the same sound – the average female formant pattern is said to be scaled upward in frequency by around 20% compared to the average male formant pattern; furthermore, a scaling factor that relates values of the formants for females and males is inversely proportional to the overall vocal tract length;
- overall spectral shape and tilt – it has been shown that male vowels have narrower formant bandwidths and less steeply sloping spectrum (around $-12dB/octave$) compared to females;
- typical power of sound – women speak with a slightly weaker voice than men (around $2dB$);

and other different attributes, such as mean airflow, glottal efficiency or amplitude of vibration that includes another scale factor of 1.2 that relates to overall larynx size [63]. The reader may refer to detailed surveys of differences between female and male voice in [10, 69].

Signal Preprocessing. First, the digital audio signal is pre-emphasized using first order all-zero filter with the coefficient in the range of $(0, 1)$. The goal of this operation is to flatten spectrum of the signal to ensure the same amount of information in the lower and higher parts of the characteristic. Next, the audio stream is divided into short frames of signal. Due to the fact that the speech signal is a non-stationary process, the standard signal processing tools (like e.g. Fourier transform) can not be directly applied to the signal. However, since the shape of the vocal tract changes slowly compared to the pitch period, it is reasonable to assume a fixed characteristic of the filter in the speech production model (the signal is considered as quasi-stationary) over a time interval on the order of $10-20ms$ [26, p. 159]. The signal is then divided into short segments using a bank of overlapping windows. In order to minimize the effect of the discontinuities at the edges of each segment in the spectral domain, the Hamming window is typically used [27, p. 231]:

$$w(n) = \begin{cases} 0.54 - 0.46cos(\frac{2\pi(n-1)}{N-1}) & \text{for } n \in (0, N-1) \\ 0 & \text{otherwise} \end{cases} . \tag{2.1}$$

Finally, the speech/non-speech or voiced/unvoiced detection is applied. The speech/non-speech segmentation is obtained by first training a Gaussian Mixture Model (GMM) with two mixtures. The mixture with largest energy coefficient is labelled as speech and the other as silence and then followed by the classification of the frames. For the voiced/unvoiced detection, we use *Robust Algorithm for Pitch Tracking* (RAPT) method based on cross-correlation and dynamic programing technique which is described in Section 2.2.1.

2.2.1 Voice Source Related Features

It is a commonly known phenomenon that female and male voices differ in the value of pitch frequency. As mentioned in Section 2.2, the typical value of pitch frequency for females ranges between 170-275Hz and for males between 112-146Hz under clean conditions [69]. This fact was utilized in the early research in audio-based gender recognition which used the value of pitch frequency as a feature. The discriminative role of the speaker pitch frequency of voicing was experimentally confirmed for the clean speech data collected from 52 speakers in [10]. In further studies, the pitch frequency value was combined with the information provided by the acoustic analysis to identify the gender of a subject on the telephone speech or artificially corrupted data [46, 58, 71]. However, a very limited number of works was devoted to analyze the suitability of the pitch frequency to AGR problem in realistic scenarios.

There exist two main practical problems when using the pitch frequency of voicing as a parameter. First, the reliability of pitch frequency estimation can be easily affected by existence of low-frequency noise in recordings or any other degradation of speech quality. Second, the value of pitch frequency changes with the physical and emotional state of a subject. Humans, while speaking spontaneously, often raise their pitch in order to stress some parts of utterance or to make their voices more audible in the presence of high level background noise. Thus, the values of pitch frequency obtained under real conditions may highly deviate from those pre-allocated to females and males under clean conditions and, as a consequence, performance of AGR system may get affected in realistic scenarios.

In order to estimate the values of a pitch frequency the *Robust Algorithm for Pitch Tracking* ($RAPT$) was used [60][3]. The RAPT algorithm consists of the two main steps: (a) each speech frame is assessed to be voiced or unvoiced, and then, (b) the pitch frequency is estimated only for the voiced frames. The method is based on both the normalized cross-correlation which is used to obtain a set of the pitch value candidates, and dynamic programming which is employed to find the optimal pitch track. The RAPT is a widely used pitch frequency estimation algorithm thanks to its robustness and computational efficiency. The robustness of the method originates from the fact that to determine optimal voicing state and pitch frequency value, it uses not only the local information about periodicity but also estimates provided by adjacent frames. The details about the RAPT method can be found in [60]. The experimental evaluation of the usability of the pitch frequency to the AGR problem under varying conditions is presented in Chapter 3.

2.2.2 Vocal Tract Related Features

As described in Section 2.2, in addition to the differences in the value of pitch frequency, the female and male voices differ in the entire range of their spectral

[3]Precisely, the ESPS implementation of the RAPT method available in the Tcl/Tk SNACK library was used [57].

characteristics due to dissimilarities in the anatomical structure of the vocal tracts. Thus, the goal is to find features describing the shape of short-term speech spectrum, i.e. ensure good discrimination between genders and robustness under varying conditions. Since differences in the formant characteristics (frequency, bandwidth, amplitude) can be observed for females and males, one of the solutions involves the usage of these parameters to distinguish the gender. The discussion about application of the formant characteristics to AGR problem is given in Section 2.2.2.1. Another type of features that capture information encoded in the shape of the spectrum of speech signal is the parametric represention. Three different types of the parametric *cepstral* features used in the state-of-the-art Automatic Speech and Speaker Recognition systems are discussed, namely the *Linear Prediction Cepstral Coefficients* (*LPCCs*) [27, p. 309], *Mel-Frequency Cepstral Coefficients* (*MFCC*) [14] and *Perceptual Linear Prediction* (*PLP*) coefficients [23].

2.2.2.1 Formants and Bandwidths

The shape of the spectral characteristic of speech signal can be directly described by the values of formant frequencies and bandwidths. The comparison of F0 and formant features (F1-F4 with their respective frequency, amplitude and bandwidth) for ten vowels was presented in [10]. The studies on clean-condition speech data revealed that first four formant frequencies are superior to corresponding formant amplitudes and bandwidths, and that formant frequencies are slightly better than F0. When using values of the formants and bandwidths as features in realistic scenarios, the two similar problems might arise as while using the pitch frequency as a feature (see discussion in Section 2.2.1). The first problem concerns with reliable estimation of the parameters. First, estimation of formant peaks may be difficult to localize due to co-existence of other harmonics, noise, or simply degradation of recording quality. Second, these parameters can be sensitive to the physical and emotional state of a subject, and the acoustic properties of an environment in which the subject stays, e.g. background noise level.

In order to estimate the values of formant frequencies and bandwidths the method implemented in the Snack library was used [57]. In this method, the formant trajectories are determined using the dynamic programming with constraints subject to frequency continuity. The formant frequencies candidates are determined from the roots of the linear prediction polynomial function that is computed periodically. A modified version of the Viterbi algorithm is used to minimize the total cost of connecting all the mappings of the complex roots to formant frequencies between consecutive time frame. More information about the method can be found in [2]. The experimental evaluation of the formant features to the AGR problem under varying conditions is presented in Chapter 3.

2.2.2.2 Parametric Representation

As explained in Section 2.2.2, the goal is to find features describing the shape of short-term speech spectrum, i.e. ensure good discrimination between genders and robustness under varying conditions. The parametric representation of the speech signal allows to describe the shape of the spectrum using the set of mathematically derived variables. According to the speech production model presented in Section 2.2, the voiced signal is produced as the convolution of the excitation waveform generated by the glottis and the impulse response of the vocal tract. The aim is to decompose the speech signal back into these two components. Let the excitation signal $e(n)$ be convolved with the transfer function of the vocal tract filter $h(n)$ in the time domain:

$$s(n) = e(n) * h(n) \tag{2.2}$$

Then, the spectrum of the speech signal can be seen as a multiplication of the spectra of those two components:

$$|S(\omega)| = |E(\omega)| \cdot |H(\omega)| \tag{2.3}$$

where $|E(\omega)|$ encodes the fast variations (fine structure) and $|H(\omega)|$ the slow variations (envelope) of $|S(\omega)|$. The vocal tract excitation $|E(\omega)|$ features are voicing, amplitude and pitch frequency, whereas the spectral envelope $|H(\omega)|$ embodies the vocal tract resonances with their locations and bandwidths [19].

The spectral envelope can be characterized by *cepstrum* or by *linear prediction* (LP) parameters and their transformations. Let us first focus on the homomorphic speech processing which enables the separation of signals which are composed through multiplication or convolution, and then return to the linear prediction approach while discussing LPCC features. The homomorphic analysis deconvolves the vocal tract response $h(n)$ from the excitation signal $e(n)$. The multiplication of the two magnitude aforementioned spectra is converted to addition by logarithm operation. Once the components are additive, they can be separated more easily using the filtering techniques.

$$log(|S(\omega)|) = log(|E(\omega)|) + log(|H(\omega)|) \tag{2.4}$$

The *cepstrum* (anagram of the word spectrum) of the speech signal is then computed as:

$$IDFT(log(|S(\omega)|) = IDFT((log|E(\omega)|)) + IDFT(log(|H(\omega)|)), \tag{2.5}$$

where $IDFT$ denotes the inverse discrete Fourier transform. This method allows to approximately separate features characteristic for the slow varying component, $\hat{h}(n)$, and the fast varying component, $\hat{e}(n)$, from each other:

$$c(n) = \hat{s}(n) = \hat{e}(n) + \hat{h}(n). \tag{2.6}$$

Analogically to the DFT coeffcients, the low-order cepstral coeffcients represent the proprieties of the slow varying component $\hat{h}(n)$ i.e. the temporary shape of the vocal tract; the high-order cepstral coeffcients describe the fast varying component $\hat{e}(n)$ i.e. the properties of the excitation source. Typically, the set of first 8 to 14 values of $c(n)$ is assumed to correspond to $\hat{h}(n)$.

In this work we focus on the standard parametric cepstral features used in the state-of-the-art Automatic Speech and Speaker Recognition systems. First, the *Linear Prediction Cepstral Coefficients* (*LPCCs*) [27, p. 309] extracted based on the *linear prediction* (*LP*) analysis are presented. Then, the two perceptually motiveted features are compared, namely the *Mel-Frequency Cepstral Coefficients* (*MFCC*) [14] and *Perceptual Linear Prediction* (*PLP*) coefficients [23].

Linear Prediction Cepstral Coeffcients (LPCCs). As mentioned before, the spectral envelope can be also characterized by the *linear prediction* (*LP*) parameters and their transformations. In linear prediction analysis, the vocal tract transfer function is modelled by an all-pole filter with transfer function [27, p. 290]:

$$H(z) = \frac{1}{\sum_{i=0}^{p} a(i) z^{-i}}, \tag{2.7}$$

where p is the order of linear prediction and $a(0) = 1$. Taking inverse z-transform of Equation 2.7 results in:

$$s(n) = \sum_{i=1}^{p} a(i) s(n-i) + e(n). \tag{2.8}$$

The current sample of the signal is estimated as a weighted linear combination of its past p samples:

$$\hat{s}(n) = \sum_{i=1}^{p} a(i) s(n-i) \tag{2.9}$$

and the prediction error when using this approximation is:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} a(i) s(n-i). \tag{2.10}$$

The filter coeffcients $a(i)|_{i=1,\ldots,p}$ are estimated so as to minimise the mean square filter prediction error summed over the analysis window. The *autocorrelation method* can be used to determine LP filter coeffcients $a(i)|_{i=1,\ldots,p}$ [27, p. 295]. The n^{th} linear prediction cepstral coefficient (LPCCs) is then computed using a simple recursion [27, p. 310]:

$$c(n) = \begin{cases} -a(n) - \frac{1}{n} \sum_{i=1}^{n-1} (n-i)\, a(i) c(n-i) & \text{for } 0 < n \leq p \\ -\frac{1}{n} \sum_{i=n-p}^{n-1} (n-i)\, a(i) c(n-i) & \text{for } n > p \end{cases}. \tag{2.11}$$

The principal advantage of the cepstral coefficients is that they are approximately decorrelated. However, the problem is that the high-order coefficients are numerically quite small when comparing to the low-order coefficients, and this results in a very wide range variances when going from the low to high cepstral coefficients. However, this problem can be alleviated by introducing the so called *cepstral liftering*, namely the process of re-scaling the cepstral coefficients to have similar magnitudes [70, p. 64].

Mel Frequency Cepstral Coefficients (MFCCs). Let us first introduce the process of extraction the *Mel-Frequency Cepstral Coefficients* (*MFCC*) [14]. In order to estimate MFCCs, first the power spectrum of the signal is estimated using the *Short Time Fourier* transform. In other words, the short-term spectrum is obtained by appling the FFT algorithm to $10 - 20ms$ frame of the windowed signal. Next, the frequency axis of the spectrum is rescaled in accordance with the studies on human speech perception. It has been shown that the relation between the objective frequency (expressed in *Hertz*) and the subjective frequency perceived by humans is almost linear up to $1kHz$, and logarithmic above this value (see Figure 2.3). Davis and Mermelstein [14] observed that transformation of the frequency axis of the spectrum from linear to *mel scale* (perceptual scale of pitches judged by listeners) highlights the perceptually relevant properties of the spectrum. To rescale the frequency axis, the short-term speech spectrum is processed in the bank of overlapping mel-frequency filters presented in Figure 2.4. This way, the FFT magnitude coefficients are weighted by transfer functions of the triangular filters and logarithm operation is performed to produce logarithmic filterbank energy coefficients, $m_{j|j=1,\ldots,P}$, where P is the number of filters in the bank. The cepstrum of the signal is then obtained by IDFT. Owing to symmetry of the log magnitude spectrum function, a Fourier transform can be replaced by the less computationally demanding Discrete Cosine transform (DCT). Finally, the n^{th} mel-frequency cepstral coefficient (MFCCs) is computed using the following formula [26, p. 163]:

$$c(n) = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} A_j cos\left(\frac{\pi n\,(j-0.5)}{P}\right) \qquad \text{for } n \in [1, P], \tag{2.12}$$

where A_j is equal to the logarithm of the magnitude coefficient m_j.

Perceptual Linear Prediction (PLP) coefficients. PLPs versus MFCCs. Let us discuss the extraction process of he *Perceptual Linear Prediction* (*PLP*) [23] coefficients, and study their properties with respect to MFCCs based on [19, 26].

Apart from the last step of analysis, the extraction process of PLP features is very similar to the extraction process of MFCCs, but with perceptual properties incorporated in a way that it is more directly related to psychophysical results. Similarly to MFCCs, first the short-term power spectrum is estimated, and then, the perceptual properties are introduced when the signal is processed by the filter bank. However, instead of using the triangular mel filters, the trapezoidally

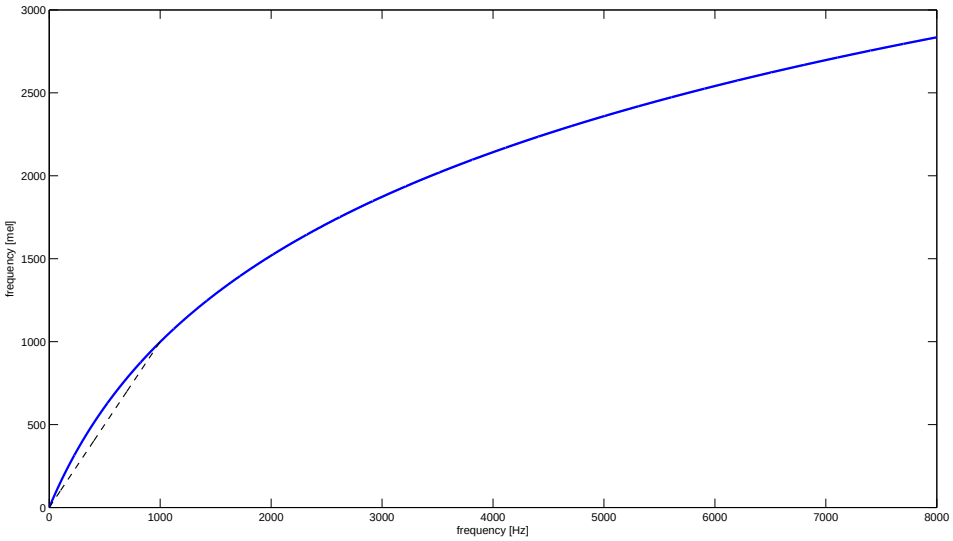


Figure 2.3. Relation between the objective frequency [Hz] and the frequency perceived by humans according to the *mel scale*: $mel(f) = 1125 \cdot ln(1 + \frac{f}{700})$ [27, p. 34]. Characteristic is almost linear below $1kHz$.

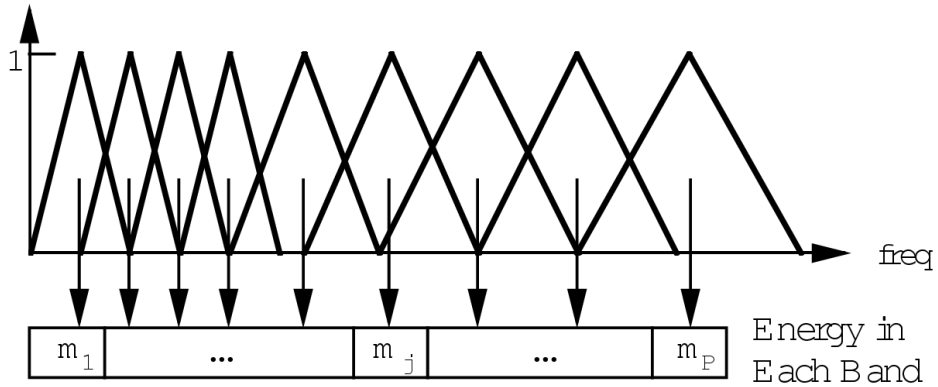


Figure 2.4. Triangular filters of the type suggested by Davis and Mermelstein [14] for transforming the frequency axis of the short-term speech spectrum onto the mel-scale.

shaped filters are applied at roughly 1-Bark intervals imitating the critical-band filters. Then, the spectrum is pre-emphasized by a function that approximates the sensitivity of human hearing at different frequencies, unlikely in the case of MFCCs, where the preemphasis is done before the log power spectrum estimation. The pre-emphasis is made by weighting the elements of the critical band spectrum. Next, the compression of the spectral amplitudes is performed. The weighted elements of the critical band spectrum are compressed by cubic root operation to approximate the non-linear relationship between the intensity of a sound and its perceived loudness. Next, the Inverse Discrete Fourier Transform (IDFT) is taken. In case of

MFCCs, this step provides the cepstral coefficients which are approximately orthogonal, while for PLP analysis the results are more like autocorrelation coefficients. Finally, a lower order all-pole model of LPC is applied to perform spectral smoothing, and in consequence, provide compact approximation of the spectrum. The LP parameters are converted to cepstral coefficients through the simple recursion (as given in Equation 2.11). This way the orthogonal representation of the features is obtained. Summarizing, the principal difference between the MFCC and PLP features lies in the nature of spectral smoothing. In case of MFCCs, the spectral smoothing is based on the cepstral analysis, whereas for PLPs, it is based on the linear prediction analysis. The experimental evidence suggested that in overall ASR systems based on PLPs provide comparable performance as those based on MFCCs [26, p. 165].

The three discussed cepstral representations have a similar structure of a feature vector. The feature vector $\overline{c}_t$ extracted from a short frame of speech signal at time t, contains a set of the *static* coefficients $c(n)|_{n=1,..,N_c-1}$. N_c denotes the number of the initial coefficient of $c(n)$ included to the feature vector and is typically chosen between 8 and 14. Additionally, the *energy* coefficient $c(0)$ is typically added to the static feature vector such that $\overline{c}_t = [c(0), .., c(N_c - 1)]$. Furthermore, in order to take into account the time correlation that exist in the speech signal due to coarticulation, the time derivatives may be added to the static parameters. The first order regression coefficients referred to as *delta* coefficients ($\Delta\overline{c}_t$) are considered [27, p. 425]:

$$\Delta\overline{c}_t = \overline{c}_{t+D} - \overline{c}_{t-D}, \tag{2.13}$$

where D represents the number of frames to offset either side of the current frame and is typically set to value of 2. However, since the time-difference features are usually sensitive to random fluctuations in the original static features, a more robust measure of local change is obtained by applying linear regression over a sequence of frames [26, p. 166]:

$$\Delta\overline{c}_t = \frac{\sum_{\tau=1}^{D} \tau \left(\overline{c}_{t+\tau} - \overline{c}_{t-\tau}\right)}{2\sum_{\tau=1}^{D} \tau^2} \tag{2.14}$$

The delta coefficients are determined for all the static parameters including the energy coefficient. Thus, the total number of the elements in the feature vector is $2N_c$ when adding the delta coefficients.

In the literature, several studies have been reported using different parametric representations of speech signal to AGR problem. The analysis of the linear prediction, autocorrelation, reflection and cepstral representations was performed on the clean speech data collected from 52 speakers for vowels, voiced and unvoiced fricatives [69]. It was found that cepstral features yield the best system and the performance improves when increasing linear prediction order from 8 to 20. It was also observed that AGR system for vowels and voiced fricatives attain better performance than for unvoiced fricatives. In addition, the study implied that gender

information is time invariant, phoneme independent, and speaker independent for a given gender. The evaluation of the 9 initial MFCCs for different groups of phonemes was conducted in [18]. The study showed that AGR based on vowels, nasal, liquids perform better than AGR based on fricatives, stops, and silence and sound 'H'. It was also found that the static coefficients are superior to the delta coefficients, and that using of both types of coefficients (static+delta) may improve performance. More recently, the comparison of Support Vector Machines (SVMs) with nearest neighbor classifiers for the first 12 cepstral coefficients was presented in [68]. For high quality recordings taken from the ISOLET corpus, the system based on SVMs attained perfect recognition rate. Summarizing, the aforementioned studies were conducted on high-quality, clean-condition speech data. However, a very limited number of works considered performance of the features in realistic scenarios. Similar to the case as formants and bandwidths, the cepstral features can be affected in realistic scenarios through for example, environmental variations and background noise, poor quality microphone or speaker varying intensities. This motivated us to revisit AGR studies on the cepstral features. We present this study in Chapter 3.

2.3 Visual Features

A face is often regarded by humans as one of the most important cue in determining gender of a person. Numerous studies have been conducted to indicate key features of the face that allow humans for very precise and fast recognition of gender [6, 42]. Psychologists aimed to determine exact parts or attributes of the face that are essential for this process. The eye and brow region was identified as important when a straight view of the face was considered [8], but nose and chin protuberance in a three-quarter view [6, 11]. Further studies on sexual dimorphism in the human face showed that the importance and typical qualities of different characteristics heavily depend on age and other individual attributes of a subject [54]. Thus, indication of a set of features that unambiguously describes a subject is open to research. Additionally, the selection of adequate features for robust AGR is further impeded by factors that are challenging for all vision-based systems, like quality and scale of an image, orientation and alignment of an object on an image or changing lightning conditions.

Early research in AGR have focussed upon the use of artificial neural networks (ANNs) to both extract relevant features from raw images and perform classification. These systems were evaluated on very small sets of subjects and clean condition data. Perfect performance was reported for the experiments conducted on a set of 20 subjects using the 2 layer back-propagation ANN called "*Empath*" [13]. Similarly, the average accuracy of 91.1% was obtained for 3 layer back-propagation ANN called "*SEXNET*" that classified a set of 90 exemplars [20]. Subsequently, the applicability of geometrical features indicated by psychological researches was verified. The 16 geometrical features (e.g. eyebrow thickness or nose width) were extracted from frontal view images of 42 subjects and used for experiments with two

competing hyper basis function networks corresponding to each gender (one for male recognition and one for female recognition). The accuracy of 79% was obtained for clean condition data [7]. Unfortunately, this approach did not provide satisfactory results and the process of feature extraction was found to be time-consuming and complex.

The standard *Automatic Face Recognition (AFR)* systems use a low-dimensional representation of faces obtained by means of component analysis techniques. The eigenface method exploits *Principal Component Analysis (PCA)* to extract the most relevant characteristics of human faces. The suitability of the method for the V-AGR problem and clean condition data was experimentally confirmed by the comparison with raw data representation [68]. However, in case of face recognition systems, this technique works efficiently only when constant face pose and lightning are preserved and tends to fail under varying conditions. To overcome this problem a technique that additionally uses *Linear Discriminant Analysis (LDA)*, referred to as the fisherface method, was introduced [3]. Details of the eigenface and fisherface techniques are presented in the following sections.

2.3.1 Eigenfaces

The eigenface method developed by Sirovich and Kirby [56] and used by Turk and Pentland [64] exploits statistical analysis to extract the most relevant characteristics of the human face. A set of training images is analyzed using PCA in order to find the most significant "ingredients" of the faces. Each of these ingredients is referred to as the *eigenface* [64]. Given an estimate of eigenfaces, any human face can be represented as weighted linear combination of eigenfaces. For instance, an arbitrary face can be represented as a sum of the average face plus 60% of the first eigenface, 17% of the second eigenface, and so on. The idea behind the eigenface method is to use a linear transformation to project face images into a new lower dimensional discriminant space (feature) space.

Let us consider a training set consisting of M face images, where each image is of size $N \times N$ pixels and is aligned as a column vector $\Gamma_{j|j=1,\ldots,M}$ of size $N^2 \times 1$. A linear projection of an image into a new P-th dimensional discriminant space ($P < M$) is defined as:

$$\Omega_j = \mathbf{W}^T \Gamma_j, \qquad \text{where } j = 1, \ldots, M \tag{2.15}$$

and $\mathbf{W}_{(N^2 \times P)}$ denotes a projection matrix with orthonormal columns. The image is represented in the discriminant space as $\Omega_j = [\omega_{1j}, \omega_{2j}, \ldots, \omega_{Pj}]$, where each element $\omega_{kj|k=1,\ldots,P}$ describes the contribution of the *k-th* eigenface in representing the face image Γ_j.

The optimal projection matrix $\mathbf{W}_{opt}$ is chosen to maximize the determinant of the total scatter (covariance) matrix $\mathbf{S}_T$ of the projected images, i.e. a set of orthonormal vectors $u_{n|n=1,\ldots,M}$ which will best describe the distribution of the

data that is in the scope of interest [3]:

$$\mathbf{W}_{opt} = \arg\max_{W} |\mathbf{W}^T \mathbf{S}_T \mathbf{W}| = [u_1, u_2, \ldots, u_M]. \tag{2.16}$$

The total scatter matrix $\mathbf{S}_T$ is defined as:

$$\mathbf{S}_T = \frac{1}{M} \sum_{n=1}^{M} (\Gamma_n - \Upsilon)(\Gamma_n - \Upsilon)^T = \mathbf{A}\mathbf{A}^T, \tag{2.17}$$

$$\Upsilon = \frac{1}{M} \sum_{n=1}^{M} \Gamma_n, \tag{2.18}$$

$$\mathbf{A} = [\Gamma_1 - \Upsilon, \Gamma_2 - \Upsilon, \ldots, \Gamma_M - \Upsilon], \tag{2.19}$$

where Υ denotes the average face vector over all training images and $\mathbf{A}$ consists of a set of normalized face images. The set of normalized face images is analysed using PCA technique in order to find a set of orthonormal vectors $u_{n|n=1,\ldots,M}$. However, since the matrix $\mathbf{S}_T$ is of size $N^2 \times N^2$, the application of PCA for typical image size can become a computationally expensive task (e.g. for $N = 64$, $\mathbf{S}_T$ is 4096×4096). To overcome this problem, the eigenvalues and eigenvectors of $\mathbf{S}_T$ are estimated in accordance with the assumption that the total number of training images is smaller than the dimension of a new discriminant space ($M < N^2$). Then there is only $M - 1$, rather than N^2 meaningful eigenvectors (the remaining eigenvectors will have associated eigenvalues equal to zero) [64]. Therefore, it is possible to first compute the M eigenvectors $v_{n|n=1,\ldots,M}$ of matrix $\mathbf{L}_{(M \times M)}$ given as:

$$\mathbf{L} = \mathbf{A}^T \mathbf{A}, \qquad \text{where } L_{ml} = (\Gamma_m - \Upsilon)^T (\Gamma_l - \Upsilon) \tag{2.20}$$

and then, expolit them to determine the eigenvectors of $\mathbf{S}_T$:

$$u_n = \sum_{k=1}^{M} v_{nk} (\Gamma_k - \Upsilon), \qquad \text{where } n = 1, \ldots, M. \tag{2.21}$$

Usually, only an arbitrary number of eigenvectors P (out of $M-1$), which correspond to the highest eigenvalues, is kept constituting a set of the *eigenfaces*. As discussed earlier, when $\mathbf{W}_{PCA} = \mathbf{W}_{opt} = [u_1, u_2, \ldots, u_P]$, the projection of an arbitrary image Γ_j to the eigenspace is given in the form of a vector Ω_j, such as:

$$\Omega_j = \mathbf{W}_{PCA}^T (\Gamma_j - \Upsilon). \tag{2.22}$$

The eigenface algorithm is considered as a computationally efficient and fast feature extraction method, espacially when comparing with techniques based on metric representation of the human face. The relatively high recognition rates have been obtained both for AFR and AGR task when images of identical iluminantion and resolution, and possibly of the same alignment, rotation and pose of the face were analyzed [64, 68]. The main drawback of the method is its high sensitivity to variations of the aforementioned factors. In case of automatic face recognition, the performance of the algorithm was significantly reduced for horizontal and vertical misalignments, iluminantion changes and low resolution images [33].

2.3.2 Fisherfaces

The eigenface algorithm aims at indicating the most significant variations in the set of analyzed face images in order to capture subject specific features. However, in most of the cases, the variations introduced by the alternating external conditions (e.g. illumination) and image setup (e.g. face alignment and pose) are much more significant then those introduced by individual subjects characteristics [41]. Therefore, when using the eigenfaces, the subject related variations may be "concealed" by wider scatter of the data in the direction of the varying external conditions and image setup. To overcome this problem, a method that maximizes scatter between images belonging to different classes while minimizing the scatter between images belonging to the same class was introduced and is referred to as the fisherface method [3]. This method uses the Fisher's *Linear Discriminant Analysis* (*LDA*) and can be considered as an enhanced version of the eigenface algorithm.

Let us consider a pattern recognition problem with C different classes denoted as $\chi_1, \chi_2, \ldots, \chi_C$, where each of the classes $\chi_{i|i=1,2,\ldots,C}$ contains K images $\Gamma_{j|j=1,2,\ldots,K}$. When the mean image Υ_i for each class χ_i is computed as [3]:

$$\Upsilon_i = \frac{1}{K}\sum_{j=1}^{K}\Gamma_j \tag{2.23}$$

and the mean over all images Υ using Equation 2.18, the within-class scatter matrix S_W and the between-class scatter matrix S_B are determined as follow:

$$\mathbf{S}_W = \sum_{i=1}^{C}\sum_{\Gamma_j\in\chi_i}(\Gamma_j - \Upsilon_i)(\Gamma_j - \Upsilon_i)^T \tag{2.24}$$

$$\mathbf{S}_B = \sum_{i=1}^{C}|\chi_i|(\Upsilon_i - \Upsilon)(\Upsilon_i - \Upsilon)^T \tag{2.25}$$

where $|\chi_i|$ denotes a number of samples in a class χ_i. The optimal projection matrix, $\mathbf{W}_{opt}$, should be chosen to miximize the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix, i.e. [3]

$$\mathbf{W}_{opt} = \arg\max_{W}\frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|}. \tag{2.26}$$

The optimal projection matrix consists of generalized eigenvectors of $\mathbf{S}_B$ and $\mathbf{S}_W$ and to be determined requires calculation of $\mathbf{S}_W^{-1}$. However, due to the fact that the rank of the within-class scatter matrix $\mathbf{S}_W$ ($N^2 \times N^2$) is smaller than $M - C$ and, as assumed before, the number of images in the set $\mathbf{S}$ is usually smaller than the number of pixel in an image ($M < N^2$), the matrix $\mathbf{S}_W$ is singular. To overcome this problem, the fisherface method employs two step algorithm to obtain $\mathbf{W}_{opt}$. First, the PCA analysis is used to reduce the dimension of feature space from N^2 to $M - C$, and $\mathbf{W}_{PCA}$ is estimated as described in Section 2.3.1. Second, the Fisher's

LDA is used to obtain projection into $C-1$ subspace. The auxiliary projection matrix, $\tilde{\mathbf{W}}_{LDA}$, is determined by solving the following equation:

$$\tilde{\mathbf{S}}_B\mathbf{V} = \tilde{\mathbf{S}}_W\mathbf{V}D, \tag{2.27}$$

where $\mathbf{V}$ is a matrix contaning eigenvectors and D is a vector containing corresponding eigenvalues, and

$$\begin{aligned}\tilde{\mathbf{S}}_B &= \mathbf{W}_{PCA}^T\mathbf{S}_B\mathbf{W}_{PCA},\\ \tilde{\mathbf{S}}_W &= \mathbf{W}_{PCA}^T\mathbf{S}_W\mathbf{W}_{PCA}.\end{aligned}$$

$\tilde{\mathbf{W}}_{LDA}$ is created as a submatrix of $\mathbf{V}$ where only $C-1$ eigenvectors corresponding to the highest eigenvalues in D is kept. Finally, the optimal projection matrix for the fisferface method, $\mathbf{W}_{LDA}$, is determined as:

$$\mathbf{W}_{LDA} = \mathbf{W}_{PCA}^T\tilde{\mathbf{W}}_{LDA}. \tag{2.28}$$

The projection of data into the fisherspace is achieved analogously to the projection of the data into the eigenspace, i.e. an arbitrary image Γ_j is to be represented in the fisherspace as a vector $\tilde{\Omega}_j$, where:

$$\tilde{\Omega}_j = \mathbf{W}_{LDA}^T(\Gamma_j - \Upsilon). \tag{2.29}$$

In case of the large variation in lightning and facial experssions for the AFR problem, the fisherface method has been shown to be superior to the eigenface method [3]. However, the eigenfaces can outperform fisherface when the number of samples (images) per class is small [36].

Chapter 4 presents experimental studies using eigenfaces and fisherfaces under varying conditions.

2.4 Classification

The goal of classification is to predict a class label of an object given the feature attributes. Each of the classes contains a group of items of certain properties which are described by a model (see Section 2.1). Depending on the method used to represent the model, the classification algorithms can be roughly divided into the two following categories [5, p. 43][43]:

1. the *generative classifiers* which determine the class-conditional densities (*likelihood)* $P(\mathbf{x}|y_i)$ of the input vector $\mathbf{x}$ for each class y_i and infer the prior class probabilities $P(y_i)$ to calculate the posterior class probabilities $P(y_i|\mathbf{x})$ using Bayes' rule[4]:
$$P(y_i|\mathbf{x}) = \frac{P(\mathbf{x}|y_i)P(y_i)}{P(\mathbf{x})} \tag{2.30}$$
and then picking the most likely class;

[4]Equivalently, the joint probability $P(\mathbf{x}, y_i)$ can be model directly and then normalized to obtain the posterior probabilities.

2. the *discriminative classifiers* which form *discriminant functions* that directly maps input $\mathbf{x}$ into decision (class label y_i) under the constraint of minimizing classification error.

From the cue integration point of view, the main advantage of the generative classifiers consist in the possibility of combining the outputs of a few classifiers systematically by using the rules of probability, and is grounded on the fact that each of the models give estimates of posterior probability for the classes. However, at the same time, the process of modelling the probability densities of appropriate accuracy requires large amount of training data, especially when input is of high dimensionality. In contrast, the discriminative classifiers by direct forming of discriminant functions avoid spending computational resources on modelling the probability distributions. However, additional methods are required in order to combine the outputs of such classifiers in a systematic way, such as taking into account confidence with which a sample is assigned to a particular class (see Section 2.5).

In this work we use the *Support Vector Machines* (*SVMs*) which are an example of a discriminant classifier. The SVMs were chosen as classification algorithm due largely to their superior performance compared to other classifiers in the previous studies on AGR problem, both for audio and visual data. The comparison of SVMs with nearest neighbor classifiers for the first 12 cepstral coefficients on high quality audio recordings from the ISOLET corpus was presented by Walawalkar *et. al* [68] with 100% AGR rate for SVMs. Similarly, for the AGR task based on visual cues the experimental studies conducted by Walawalkar *et. al* [68] and Moghaddam *et. al* [39] suggested that the SVMs with the *Radial Basis Function (RBF)* kernel are superior to the linear, quadratic, fisher linear discriminant, k-nearest neighbor classifiers as well as to more complex techniques such as large ensemble RBF networks. Further, the SVMs attracted our attention due to the fact of being a binary classifier and requiring a relatively small number of training examples to estimate with respect to the dimensionality of the input, unlike for instance the Multi Layer Perceptron (MLP) [24], which is especially valuable for the visual-based AGR system where we train with only a very few number of images. Important was also the issue of using the same classification algorithm both in case of A-AGR and V-AGR system.

2.4.1 Support Vector Machine

This section contains a brief description of the *Support Vector Machine* (*SVM*) which is a discriminant classifier originally proposed by Vapnik [65]. The SVMs were designed to solve a two-class problem, i.e. they are an example of a *binary* classification algorithm. Given a set of training samples $(\mathbf{x}_i, y_i)|_{i=1,\ldots,l}$, where $\mathbf{x}_i \in \mathbb{R}^N$ is a feature vector, and $y_i \in \{-1, +1\}$ is the corresponding class label, a goal is to find a *discriminant function* $f : \mathbb{R}^N \rightarrow \mathbb{R}$ that will separate the samples belonging to the two different classes. When making an assumption that samples are linearly separable, a decision surface that does the separation is given in a form

of a hyperplane [22, p. 319]:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0, \tag{2.31}$$

where $\mathbf{w}$ is an adjustable weight vector, b is a bias, and $f(\mathbf{x}) \geq 0$ for $y_i = +1$ and $f(\mathbf{x}) < 0$ for $y_i = -1$. For a given weight vector $\mathbf{w}$ and bias b, the separation between the hyperplane and the closest sample is referred to as the *margin of separation*. There exist many realizations of the hyperplane that might classify the data, however the aim of SVMs is to find a particular hyperplane for which the margin of separation is maximized for the samples from both classes. For this purpose, the following optimalization problem has to be solved [22, p. 322]: find the optimum values of the weight vector $\mathbf{w}_0$ and the bias b_0 such that they satisfy the constraints

$$y_i(\mathbf{w}_0^T\mathbf{x}_i + b_0) \geq 1 \qquad \text{for } i = 1, \ldots, l$$

and the weight vector $\mathbf{w}_0$ minimizes the cost function:

$$\phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w}.$$

This constrained optimization problem is solved using the Lagrange multipliers method [66]. The optimal parameters $\mathbf{w}_0$ and b_0 given by [22, p. 324]:

$$\begin{aligned} w_0 &= \sum_{i=1}^{l} \alpha_{0,i} y_i \mathbf{x}_i \\ b_0 &= 1 - \mathbf{w}_0^T\mathbf{x}^{(s)} \qquad \text{for } y^{(s)} = 1 \end{aligned}$$

are used to formulate the discriminant function defining the optimal separation hyperpalne:

$$f(\mathbf{x}) = \sum_{j=1}^{m} \alpha_{0,j} y_j \mathbf{x}_j^T\mathbf{x} + b_0, \tag{2.32}$$

where $\{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_m\} \in \mathbf{x}^{(s)}$ and $m \leq l$ are the *support vectors* and $\alpha_{0,j}$ are the corresponding Lagrange multipliers. It is important to note that the classifier is determined only by a subset of the training samples (i.e. support vectors), the corresponding Lagrange multipliers and the bias.

In case of the non-linear separable problem, when the training samples can not be discriminated using a linear hyperplane, it is possible to formulate the optimization problem in a way that the classification error over the training set is minimized [22, p. 328]: find the optimum values of the weight vector $\mathbf{w}_0$ and the bias b_0 such that they satisfy the constraints

$$\begin{aligned} y_i(\mathbf{w}_0^T\mathbf{x}_i + b_0) &\geq 1 - \xi_i \qquad \text{for } i = 1, \ldots, l \\ \xi_i &\geq 0 \qquad \text{for all } i \end{aligned}$$

and the weight vector w_0 and the *slack variables* ξ_i minimize the cost function:

$$\phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i.$$

where C is a user-specified positive parameter. The slack variables can be seen as a measure of the violation of the margin. For $0 < \xi_i < 1$, the sample crosses the boundary determined by the margin; however it is still correctly classified. The value greater than 1 means that the sample falls on the wrong side of the hyperplane. The parameter C controls the tradeoff between complexity of the classifier and the number of nonseparable samples, i.e. the amount of errors that are made. Therefore, it may be seen as a kind of a "regularization" parameter; it has influence on the generalization abilities of the classifier. As mentioned earlier, C is to be specified by a user, usually by means of experiments. The solution of the specified optimization problem is the discriminant function in a form previously specified by Equation 2.32.

So far we discussed how to find the optimal separation hyperplane in the linearly separable and the linearly non-separable case. However, the SVMs can be extended to the form of non-linear classifier by applying the so called *kernel trick*. The discrimination function is similar as in the linear cases, except that every inner product is replaced by a non-linear *kernel* function, $K(\mathbf{x}, \mathbf{y})$, that projects input vectors to a high-dimensional feature space [22, p. 330]:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{j=1}^{m}\alpha_j\varphi(\mathbf{x}_j)^T\varphi(\mathbf{x}_j) + b, \\ K(\mathbf{x}, \mathbf{y}) &= \varphi(\mathbf{x})^T\varphi(\mathbf{y}). \end{aligned}$$

Thus, the separation hyperplane is formulated in the high-dimensional feature space as:

$$f(\mathbf{x}) = \sum_{j=1}^{m}\alpha_j K(\mathbf{x}_j, \mathbf{x}) + b. \tag{2.33}$$

It has to be noted that by using kernel function, the costly determination of the explicit high-dimensional representation of the input vectors is avoided. The kernel function $K(\mathbf{x}, \mathbf{y})$ can be interpreted as a similarity measure between the vectors $\mathbf{x}$ and $\mathbf{y}$. The two widely used inner product kernels are [22, p. 333]:

- *polynomial kernel*:
$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + p)^d, \tag{2.34}$$
where the power d is specified a priori by the user. The *linear kernel* is a special type of the polynomial kernel: $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$.

- *radial-basis function (Gaussian) kernel*:
$$K(\mathbf{x}, \mathbf{y}) = e^{\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2\right)} \tag{2.35}$$
where the width σ^2 is specified a priori by the user.

Platt's Method

For the purpose of cue integration, described in Section 2.5, we need information about the *confidence* with which a sample is assigned to a particular class. One of the measures that expresses the confidence of classification is the conditional posteriori class probability $P(y_i|\mathbf{x})$ of classifying a sample $\mathbf{x}$ to a certain class y_i. However, as mentioned earlier, SVMs can not directly estimate $P(y_i|\mathbf{x})$.

The method derived by Platt [47] allows to estimate a value of the posteriori class probability for SVMs and is based on the following reasoning [53]. First, the value of the SVMs discriminative function $f(\mathbf{x})$ is interpreted as a distance of a sample to the optimal hyperplane. Then, when assuming that $P(y_i|\mathbf{x})$ is continuous in $\mathbf{x}$, it can be said that the samples placed closer to the hyperplane have a larger probability of being misclassified than samples lying further from the hyperplane. The closer the sample is to the hyperplane, the smaller change is enough to alter the decision sign. Thus, it is assumed that the a posteriori probability can be represented in terms of function of the value of $f(\mathbf{x})$. Several different functions were proposed to model this relation [21], however the Platt's modelling motivated by empirical results is in the widest use. In Platt's method, the relation between the posteriori probability and value of the discriminant function is expressed in terms of the sigmoid function, i.e. the sigmoid function maps the SVM outputs into probabilities [47]:

$$P(y_i = +1|f(\mathbf{x})) = \frac{1}{1 + e^{Af(\mathbf{x})+B}}. \tag{2.36}$$

As long as $A < 0$ the monotonically increasing function is assured. The parameters of the function, A and B, are found empirically during training by minimizing the cross-entropy error.

2.5 Cue Integration

Cue integration is directly linked to the broad group of *information fusion* techniques and consists in utilizing a combination of information from different sources[5], either to generate one representational format, or to reach a decision [55]. An important assumption, determining the usefulness of the approach, is that the information provided by the different signals are complementary.

While considering the cue integration in pattern recognition from the point of decision making, the main reasons for combining information are accuracy, robustness and efficiency. The accuracy of a pattern recognition system may be improved by providing more complete picture of a classified object. The use of multiple sensors, i.e. redundancy, can increase reliability of the provided description. Further, the separate streams of information may be distorted in a different manner providing different distributions of errors in the system. Then, by proper integration of the cues, the robustness of recognition process towards various distortions and

[5]Or from one source, but using different signal extraction or processing techniques [48].

noises may be improved. Finally, cost of implementation can be reduced by using several cheap sensors rather than a single expensive one.

The integration of information processed by the systems can be performed at different levels of system functionality, and the following two main approaches are distinguished among the data fusion techniques [29, 55]:

- fusion at *feature level* when information is combined before any use of classifiers or experts, and it is also often referred to as *low level* fusion;

- fusion at *classifier level* when integration is accomplished by an ensemble of classifiers or experts, and it is often referred to as *high level* fusion.

In case of the low level fusion, the combination of several information signals can be simply achieved by: (a) weighted summation of the extracted feature vectors if they are of the same type, e.g. data from two microphones; or (b) concatenation of the feature vectors if they are of different types, e.g. audio and visual data. In spite of mathematical simplicity, the low level fusion has certain drawbacks. First of all, the feature vectors originating from different sources must be available at the same time, which leads to the requirement of synchronous data acquisition. Moreover, the dimensionality of the final feature vector may be increased, which can lead to the *curse of dimensionality* problem [15, p. 170]. Due to these problems, the high level fusion is usually preferred for audio-visual integration [29].

For the purpose of high level fusion, information about the *confidence* with which a sample is assigned to a particular class is required. For instance, in case of SVMs a measure based on the distance of a sample to the margin is often employed. The method for estimating the a posteriori probability of a particular class based on a value of the SVMs discriminative function is presented in Section 2.4.1. There are different methods for integrating evidences, i.e. confidence measures provided by multiple classifiers. The common methods include [49, 55]:

- *majority voting* of classifiers, where each of the classifiers provides a hard decision about classes (0 or 1) based on evidence. A final consensus is reached when at least one more than half the number of classifiers agree on the same decision. For a two-class problem, the number of classifiers must be odd and greater than two (to prevent ties).

- *ranked list combination*, where each classifier provides a ranked list of class labels, with the top entry indicating the most preferred class. The final decision can be obtained, for instance, by selecting the most popular class among the n-top entries in the list.

- methods based on *algebraic combination* of classifier outputs. The total support for each class is obtained as a simple function (like sum, product, max, min or mean) of the evidences provided by individual classifiers. The final decision is made by choosing the class with the strongest support. Let us

assume that there are L different classifiers, where each classifier i provides evidence d_{ij} for each class j such that:

$$\begin{aligned} \underset{i=1,\ldots,L}{\forall} \underset{j=1,\ldots,N}{\forall} &: \quad 0 \leq d_{ij} \leq 1, \\ \underset{i=1,\ldots,L}{\forall} &: \quad \sum_{j=1}^{N} d_{ij} = 1, \end{aligned}$$

the class with the strongest support $j_0 \in \{1, \ldots, N\}$ is chosen as the following:

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\mathcal{F}(d_{1j}, \ldots, d_{Lj})}{\sum_{j=1}^{N} \mathcal{F}(d_{1j}, \ldots, d_{Lj})} \tag{2.37}$$

where $\mathcal{F}(\cdot)$ denotes an algebraic function and the denominator is formulated to normalize the values of the support for each class between 0 and 1. For instance, the final decisions for the sum, product or max rule is obtained as given by Equations 2.38, 2.39 or 2.40, respectively.

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\sum_{i=1}^{L} d_{ij}}{\sum_{j=1}^{N} \sum_{i=1}^{L} d_{ij}} \tag{2.38}$$

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\prod_{i=1}^{L} d_{ij}}{\sum_{j=1}^{N} \prod_{i=1}^{L} d_{ij}} \tag{2.39}$$

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\max_{i=1,\ldots,L} d_{ij}}{\sum_{j=1}^{N} \max_{i=1,\ldots,L} d_{ij}} \tag{2.40}$$

It can be observed that Equations 2.38-2.40 give equal importance to all classifiers. However, if a particular classifier has better discrimination abilities and provides more reliable decisions, the overall performance of the system may be further improved by increasing importance of evidences provided by this classifier during the cue integration. In practice, a higher weight is usually assigned to such a classifier. Assuming that weights for all classifiers are determined such that:

$$\underset{i=1,\ldots,L}{\forall} : 0 \leq w_i \leq 1 \qquad \wedge \qquad \sum_{i=1}^{L} w_i = 1,$$

the class with the strongest support $j_0 \in \{1, \ldots, N\}$ can be chosen as:

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\mathcal{F}(d_{1j}, \ldots, d_{Lj}; w_1, \ldots, w_L)}{\sum_{j=1}^{N} \mathcal{F}(d_{1j}, \ldots, d_{Lj}; w_1, \ldots, w_L)}. \tag{2.41}$$

As before, the final decisions for the sum, product or max rule is obtained using Equations 2.42, 2.43 or 2.44, respectively.

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\sum_{i=1}^{L} d_{ij} w_i}{\sum_{j=1}^{N} \sum_{i=1}^{L} d_{ij} w_i} \tag{2.42}$$

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\prod_{i=1}^{L} d_{ij} w_i}{\sum_{j=1}^{N} \prod_{i=1}^{L} d_{ij} w_i} \tag{2.43}$$

$$j_0 = \underset{j=1,\ldots,N}{\arg\max} \frac{\max_{i=1,\ldots,L} d_{ij} w_i}{\sum_{j=1}^{N} \max_{i=1,\ldots,L} d_{ij} w_i} \tag{2.44}$$

There are different methods for determing weights of the classifiers. Generally, these methods can be divided into the two groups with respect to the way of defining value of the weights. The first group includes methods with the *static* weights that are usually determined using an additional set of data (e.g. development set) by minimizing a total classification error. Each classifier is assigned a certain weight that stays constant for all the test samples. The second group is constituted by methods that determine the weights *dynamically*, separately for each of the classified samples at run time. The weight for each of the classifiers can be specified, for example, as inverse of entropy [38] or can be found using the Dempster-Shafer method [31].

Thus far, we assumed that a particular object (subject) is characterized by a single sample. Let us consider the scenario when the object is characterized by a sequence of consecutive samples, where each of the samples is classified separately, such as a series of consecutive images or frames of speech signal. In such a case, instead of performing the cue integration every time after the classfication of a single sample, the fusion can be accomplished in the so called *batch mode*, i.e. for all samples in a sequence together after the last sample in the sequence is classified [5, p. 240]. Assuming that the number of samples in the sequence is constant for each classifier $K_{i|i=1,\ldots,L}$ (but may differ for different classifiers), the overall evidence d_{ij}^{ALL} for the i-th classifier and the j-th class can be obtained based on all samples in a sequence $k = 1, \ldots, K_i$, where each of samples provides an evidence d_{ij}^{k}:

$$d_{ij}^{ALL} = \frac{\mathcal{F}\left(d_{ij}^{1}, \ldots, d_{ij}^{K_i}\right)}{\sum_{j=1}^{N} \mathcal{F}\left(d_{ij}^{1}, \ldots, d_{ij}^{K_i}\right)} \tag{2.45}$$

Then, the cue integration is performed according to Equation 2.41, where d_{ij} is substituted by the overall sequence evidence d_{ij}^{ALL}. The evidence d_{ij}^{ALL} can be obtained for the sum, product or max rule using Equations 2.46, 2.47 or 2.48,

respectively.

$$d_{ij}^{ALL} = \frac{\sum_{k=1}^{K_i} d_{ij}^k}{\sum_{j=1}^{N}\sum_{k=1}^{K_i} d_{ij}^k} \tag{2.46}$$

$$d_{ij}^{ALL} = \frac{\prod_{k=1}^{K_i} d_{ij}^k}{\sum_{j=1}^{N}\prod_{k=1}^{K_i} d_{ij}^k} \tag{2.47}$$

$$d_{ij}^{ALL} = \frac{\max_{k=1,\ldots,K_i} d_{ij}^k}{\sum_{j=1}^{N}\max_{k=1,\ldots,K_i} d_{ij}^k} \tag{2.48}$$

2.5.1 Audio-Visual AGR System

In most typical applications of AGR system, both audio signal and visual signal are available. Ideally, an AGR system should be able to exploit both modalities to improve accuracy and robustness of the system. Since each modality has different characteristics, audio-visual cues can provide a more comprehensive description of a subject than a single modality. Finally, integration of the cues may yield a AGR system that is resilient to the degradation of both, or even to temporal unavailability of one of the input signals.

In this work, a two-fold approach is adopted in designing the audio-visual AGR system. First, we study the two cues separately by building audio-based and vision-based AGR systems. Second, these systems are integrated to provide the final decision based on both modalities. The architecture of the AV-AGR system is presented in Figure 2.5. The AV-AGR system is created by fusing evidences from the two modalities at the high level. It is due to the fact that data synchronization at this level of the audio-visual AGR system is much easier than at the lower levels of the system.

As a measure of confidence with which a sample is assigned to a particular class, the a posteriori class probability is chosen. For SVMs, value of this probability is estimated using the Platt's method described in Section 2.4.1.

The posteriori probabilities provided by the single-cue classifiers are combined using the sum, product or max rule to provide the final decision based on both modalities. The theoretical studies conducted by Tax *et. al* [61] showed that these rules are most suitable for the two-class problem in which posterior probabilities are well estimated, i.e. without a large number of extreme posterior probability estimations, one and zero. Further, it was shown by Kuncheva that the max rule is superior to average integration and majority voting when distribution of the a posterior probabilities is uniform [32]. In our system, the approach based on the majority voting is not considered, since it requires an odd number of classifiers for a two class problem and we have only one classifier for each cue.

In case when a subject is characterized by a single sample, the cue integration can be performed directly according to Equation 2.41 where the evidence d_{ij} is substituted by the probability of assigning a particular sample to the class $J = j$ by the classier $I = i$, namely the a posteriori class probability $Pr(J = j|I = i)$ for

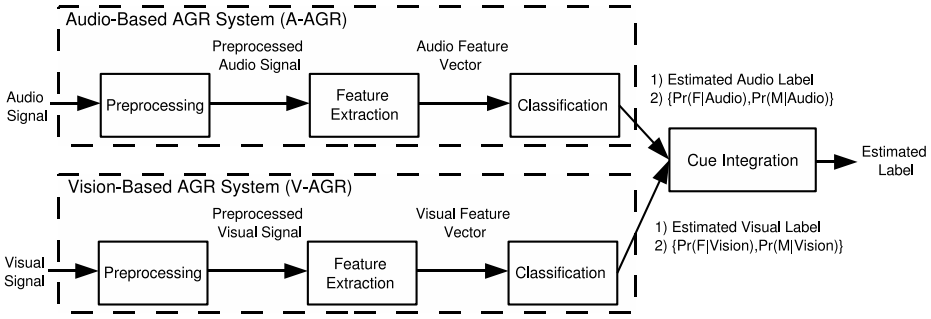


Figure 2.5. Overview of the architecture of the AV-AGR system. The two modalities are processed separately, and then integrated at the classifier level.

$j \in \{Female, Male\}$ and $i \in \{Audio, Vision\}$[6]. In case where a subject is characterized by a sequence of consecutive samples, the cue integration can be performed according to Equation 2.41 where the evidence d_{ij} is substituted by the evidence d_{ij}^{ALL} determined based on all samples in audio and visual sequences according to Equation 2.45, which takes the following form for the AV-AGR problem:

$$d_{i,j}^{ALL} = \frac{\mathcal{F}\left(Pr_1(j|i), \ldots, Pr_{K_i}(j|i)\right)}{\sum_{j=1}^{2} \mathcal{F}\left(Pr_1(j|i), \ldots, Pr_{K_i}(j|i)\right)}.$$

For example, the evidence for the female class, audio cue and sum rule, where K_A denotes the number of samples in corresponding audio sequence is obtained as[7]:

$$d_{A,F}^{ALL} = \frac{\sum_{k=1}^{K_A} Pr_k(F|A)}{\left(\sum_{k=1}^{K_A} Pr_k(F|A)\right) + \left(\sum_{k=1}^{K_A} Pr_k(M|A)\right)}.$$

The cue integration studies with equal and unequal weighting of modalities for the AV-AGR system are presented in Chapter 5.

[6]Thereafter, $Pr(J = j|I = i)$ is denoted by $Pr(j|i)$.
[7]Abbravations: 'F'=Female, 'M'=Male and 'A'=Audio.

Chapter 3

Audio-Based AGR Studies

This chapter presents experimental studies on different audio features for AGR task. The evaluation of the voice source and vocal tract related features is performed in the framework of A-AGR system described in the previous chapter. First, motivation of the studies is discussed and the main objectives are defined in Section 3.1. Then, experimental setup is specified in Section 3.2, and results obtained under varying conditions are presented and discussed in Section 3.3. Finally, the main conclusions drawn from the studies are summarized in Section 3.4.

3.1 Motivation and Objectives

In the literature, different feature representations of audio signal have been studied for the AGR task, such as fundamental frequency (F0), formants with their respective frequency, amplitude and bandwidth, and the cepstral features like LPCCs or MFCCs (see discussion in Section 2.2). Most of the previous works on the AGR mainly analyzed these features for specific phonemes or broad phonetic classes and using clean condition data. However, a very limited number of studies considered the performance of the audio features in realistic scenarios in which different practical difficulties may occur. As discussed in Sections 2.2.1 and 2.2.2, the F0 estimation process can be affected by existence of low-frequency noise in recordings or any other degradation of speech quality, and in addition, the value of F0 changes with the physical and emotional state of a subject. Similarly, the vocal tract features can be impacted in realistic scenarios by, for instance, environmental variations and background noise, poor quality microphone or speaker varying intensities. Alongside with the intention of studying audio-visual AGR, this motivated us to revisit AGR studies and analyze the voice source and vocal tract related features with emphasis on the following practical questions:

1. *What is the effect of data selection (selection of particular frames) on the AGR accuracy?* The previous studies have shown that AGR accuracy is not the same across all groups of phonemes. This information can be exploited to build a better AGR system that, ideally, will identify gender based on

a selected group of phonemes. However, in practice, this approach can be complex and requires the use of for example a phoneme recognizer before AGR. Our approach is based on the observation that typically voiced segments provide the best performance, as it was shown in [69, 18]. Thus, can selection of voiced frames lead to a better AGR system compared to the approach employing the entire speech segments?

2. *What is the effect of varying conditions on the performance of the AGR system trained with different voice source and vocal tract related features?* Which type of audio features is robust towards varying conditions?

3. *What is the effect of the cepstral feature dimension on the AGR accuracy?* In [69], it was shown that increase of linear prediction order and increase in number of cepstral coefficients lead to improvement in the AGR recognition rate. Does this trend hold also under noisy conditions?

4. *What is the effect of training data conditions on the AGR accuracy?* Is it better to train the AGR system on clean-condition data only or multi-condition (clean+noisy) data? How much the quality of data influences the gender information present in the voice source and vocal tract related features?

5. *What is the effect of fusing the voice source and vocal tract information on the AGR accuracy?* How complementary are the voice source and vocal tract related features?

3.1.1 Experiments

In order to answer the aforementioned questions, we perform three types of experiments. First, we compare two different data selection approaches: (a) where any speech part (containing both voiced and unvoiced speech) of the signal is used as a source of information about gender or (b) where only voiced speech is used instead (results in Section 3.3.1). Second, we study the effectiveness of different features for the AGR problem. We investigate performance of the voice source related feature (F0) under varying conditions (Section 3.3.2). Then, we evaluate two types of the vocal tract related features: (a) first four formants with their respective frequency and bandwidth; and (b) cepstral features: LPCCs, MFCCs and PLPs (Section 3.3.3). Finally, we study integration of the voice source (F0) with each of the two types of the vocal tract related features (Section 3.3.4).

3.2 Experimental Setup

Before presenting the results obtained for different audio features, we describe the audio part of the BANCA database exploited to evaluate performance of the A-AGR system, and provide information about data preprocessing and experimental setup.

3.2.1 Database

The A-AGR system was evaluated on an audio part of the BANCA database (English corpus) comprising datasets of varying complexity [4]. The audio data acquisition was performed using two microphones (poor-quality and good-quality) under three different types of conditions:

1. *Controlled*: good-quality microphone, clean conditions;
2. *Degraded*: poor-quality microphone, stable conditions;
3. *Adverse*: good-quality microphone, background noise, arbitrary conditions.

In order to evaluate the performance of the system under clean conditions the **0** protocol (matched controlled training and test conditions) is established, as specified in Table 3.1. We define three additional protocols: **A**, **B**, **C** each in two versions: *Deg* and *Adv* for degraded and adverse conditions, respectively. The three protocols differ with respect to the quality of data used for training, development and testing. The idea is to first use the clean conditions data for training and development, and test the system under noisy conditions (protocol **A**, mismatched training and test conditions). In protocol **B**, the training is done with clean condition data and the parameters of the system are tuned with noisy development data. Finally, in protocol **C** both clean and noisy data are used for training (multi-condition training). The BANCA database contains recordings collected from 52 subjects. In order to evaluate the system on the same number of known and unknown subjects, only a half of the subjects (26) were used for training. Then, all subjects (52) were divided into two groups consisting of 16 and 36 subjects which were used for development and testing. More information about the database and setup can be found in Appendix A.

3.2.2 Analysis of Audio Data

The audio signal was sampled at $16kHz$ and analyzed in frames of 25ms using a frame shift interval of 10ms. The informative part of the signal was obtained using a speech/non-speech or voiced/unvoiced detection. For each utterance, the speech/non-speech segmentation is obtained by first training a GMM with two mixtures. The mixture with largest energy coefficient is labelled as speech and the other as silence, and then followed by the classification of the frames. The RAPT method algorithm was used both to obtain voiced frames and F0 estimates [60] (details in Section 2.2.1). The values of F1-F4 and B1-B4 were determined using linear prediction and dynamic programming [57] (details in Section 2.2.2.1). The three cepstral features, namely LPCCs, MFCCs and PLPs, were extracted using the HTK toolkit [70], and we analyzed their performance with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector. The choice of the particular order of cepstral features was motivated by the previous studies. First, MFCCs with 9 initial coefficients were evaluated

Protocol	Set ID	Conditions		
		TRAIN	DEV	TEST
0	Con_0	**Controlled**	**Controlled**	**Controlled**
A	Deg_A	**Controlled**	**Controlled**	Degraded
	Adv_A	**Controlled**	**Controlled**	Adverse
B	Deg_B	**Controlled**	Degraded	Degraded
	Adv_B	**Controlled**	Adverse	Adverse
C	Deg_C	**Controlled**+Degraded	Degraded	Degraded
	Adv_C	**Controlled**+Adverse	Adverse	Adverse

Protocol	Item	TRAIN	DEV	TEST
0,**A**,**B**,**C**	Subjects Σ(F,M)	26(13,13)	16(8,8)	36(18,18)
	Data per File	1.5s	1.3s	1.3s
0,**A**,**B**	# Files	104	64	144
C	# Files	104+64	64	144

Table 3.1. Experimental setup for different protocols. Abbreviations and symbols: 'Σ'=Total, 'F'=Females, 'M'=Males. Details can be found in Appendix A.

under clean conditions by Fussell [18]. Then, initial 13 coefficient are commonly used in automatic speech recognition to characterize the smooth envelope of the short-term spectrum (see Section 2.2.2.2). Typically, the voice source realted information is captured between 14 and 19 coefficient. Thus, initial 19 coefficients are commonly used in automatic speaker recognition system in order to model speaker information present in the high-frequency region.

3.2.3 Classification

We employed the SVMs implemented in the LIBSVM library to perform gender classification [9] (details in Section 2.4). The RBF kernel was used in case of multi-dimensional feature vectors, and the linear kernel in case of one-dimensional feature vectors. The parameters of the SVMs (error penalty C) and the RBF kernel (variance γ) were estimated on the development set. However, in case of simple almost linearly separable problems, the choice of C is of little importance. The a posteriori class probabilities are estimated using the Platt's method [47].

3.2.4 Performance Evaluation and Cue Integration

The A-AGR system is evaluated on the test set and its performance is expressed as percentage of correct classification, i.e. *classification accuracy*. Each frame of audio signal is assumed as a separate sample and is classified independently. As a measure of confidence with which a sample is assigned to a particular class, the a posteriori class probability is chosen. We report the results at a file level based on 1.3s of speech segment or voiced speech segment extracted from each video file

($K_A = 129$). The decision about gender for a single file is obtained by summing[1] the frame values of the a posteriori probabilities for each class over the whole audio segment, and then choosing the class with the highest score.

We also investigated the integration of F0 with the vocal tract realted features: (a) at *lower level*, by concatenating the features and training a single classifier, and (b) at *higher level*, by first training two independent classifiers, and then fusing their outputs using a linear weighted combination method, such as sum, product or max rule. In case of the high level integration, the decision about gender for a single file is obtained according to Equation 2.41 where the evidence d_{ij} estimated for the i-th classifier and the j-th class based on a single sample is substituted by the evidence d_{ij}^{ALL} determined based on all samples in a sequence $k = 1, \ldots, K_A$ according to Equation 2.45 for $j \in \{F, M\}$ and $i \in \{F0, VTF\}$[2]. For example, the evidence for the female class ($j = F$), voice source features (F0) and sum rule is obtained as:

$$d_{F0,F}^{ALL} = \frac{\sum_{k=1}^{K_A} Pr_k(F|F0)}{\left(\sum_{k=1}^{K_A} Pr_k(F|F0)\right) + \left(\sum_{k=1}^{K_A} Pr_k(M|F0)\right)}.$$

The weights for each classifier were determined to maximize performance of the A-AGR system on the development set.

3.3 Results and Discussion

This section presents results obtained during experimental evaluation of the A-AGR system. The results are followed by the discussion.

3.3.1 Frame Selection

First, we report the performances for two systems that differ in frame selection method for F0 and the cepstral features in Table 3.2. In the first system, a speech/non-speech segmentation was used to select speech frames (contains both voiced speech and unvoiced speech). The values of F0 for unvoiced frames were estimated via interpolation, and among the linear, logarithmic and Fourier interpolation method, the latter provided the best performance. In the second system, only voiced speech frames were selected, and this system provided higher recognition rates for all the evaluated features, especially under noisy condition. This is consistent with observations made in the previous studies where better AGR performance have been reported on the voiced phonemes compared to unvoiced phonemes [69, 18]. In addition, this also shows that selection of voiced speech frames can make the gender recognizer robust towards mismatched training and testing conditions (protocol **A**).

[1] Preliminary experiments performed on the development set showed that summation performs better than product and max function.

[2] Abbreviations: 'F'=Female, 'M'=Male, 'F0'=Voice Source Features (F0) and 'VTF'=Vocal Tract Features.

In the rest of the work, we report results for the system employing the voiced speech frame selection.

3.3.2 Voice Source Related Features

We investigated suitability of F0 for the AGR problem under varying conditions and the obtained results are shown in Table 3.3. When the AGR system was trained exclusively on clean condition data (protocol **0** and **A**), F0 attained perfect recognition under controlled conditions, however recognition rate was highly affected under noisy test conditions. In order to analyze the results, distributions of F0 for females and males in the training set and in the three test sets (one for each conditions) are presented in Figure 3.1. The perfect recognition under controlled conditions is a consequence of almost an ideal match between F0 distributions for the training and test data. Under degraded conditions, the low frequency noise was also estimated as F0 what resulting in high degradation of performance for females and perfect recognition for males. The adverse conditions data were collected in noisy environment under which presumably subjects tend to raise their F0 in order to make their voices more audible (*Lombard effect* [34]). As a result, the distributions of F0 for both females and males were shifted towards high values, and the mismatch between data used for training and testing occurred. In this case, significant decrease in performance for males and perfect recognition for females were observed. The addition of noisy data to the training set (protocol **C**), decreased the performance under degraded conditions, since more incorrect examples of F0 were introduced to the system. This suggest that the decrease in the performance is mainly due to F0 estimation error in the degraded signal. On the other hand, the performance under adverse conditions was increased, thus indicating that the system tries to compensate for the effect of raised F0 values.

Data Type	Feature	Accuracy [%]		
		Prot. **0**	Protocol **A**	
		Con_0	Deg_A	Adv_A
Speech	F0*	99.3	95.8	91.7
	$LPCC18_\Delta$	94.4	89.6	79.2
	$MFCC19_\Delta$	97.9	91.7	80.6
	$PLP19_\Delta$	95.8	86.1	81.3
Voiced	F0	100.0	95.8	93.1
	$LPCC18_\Delta$	97.2	98.6	96.5
	$MFCC19_\Delta$	97.9	97.9	93.1
	$PLP19_\Delta$	98.6	98.6	97.2

Table 3.2. Performance of the AGR system using all speech frames (Speech) and using only voiced speech frames (Voiced) for F0, LPCCs, MFCCs, and PLPs with 19 static and delta coefficients under three types of conditions: controlled, degraded and adverse for protocol **A**. For LPCCs energy coefficient was not determined. *F0 for unvoiced frames was estimated using the Fourier interpolation method.

Feature	ID	Accuracy [%]				
		Prot. **0**	Protocol **A**		Protocol **C**	
		Con_0	Deg_A	Adv_A	Deg_C	Adv_C
F0	Females	100.0	91.7	100.0	91.7	100.0
	Males	100.0	100.0	86.1	98.6	88.9
	Total	100.0	95.8	93.1	95.1	94.4

Table 3.3. Performance of the A-AGR system for F0. The results for the protocol **B** are identical as for **A**, since the constant values of parameter C for the SVMs with the linear kernel were assumed.

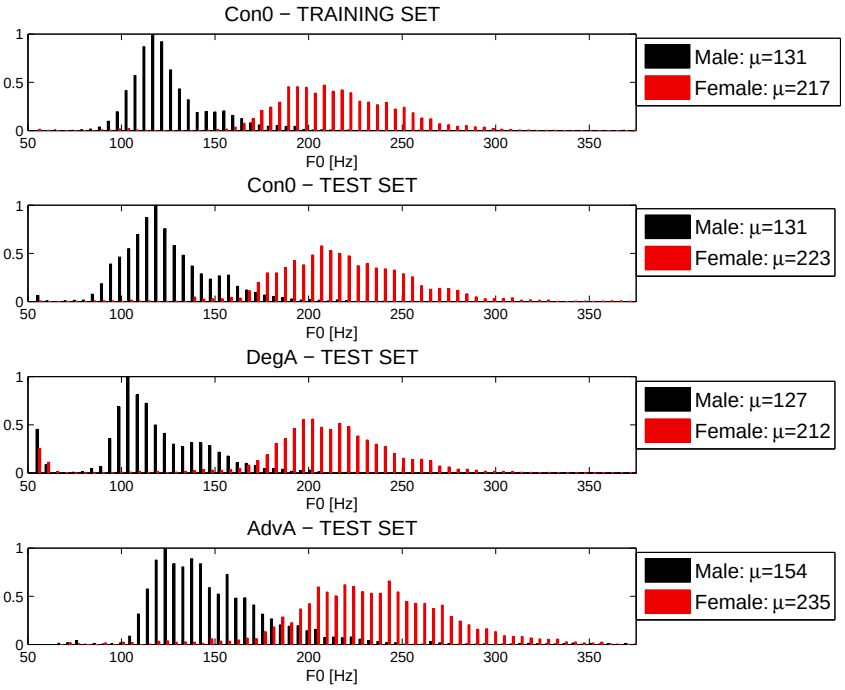


Figure 3.1. Distributions of F0 values for females and males in the training set containing controlled conditions data and the three test sets consisting of controlled (Con_0), degraded (Deg_A) and adverse (Adv_A) data.

Feature	Accuracy [%]						
	Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
	Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	100	95.8	93.1	95.8	93.1	95.1	94.4
F1-F4+B1-B4	92.4	89.6	85.4	88.2	84.7	88.2	86.1
LPCC18$_\Delta$	97.2	98.6	96.5	98.6	96.5	100	97.9
MFCC19$_\Delta$	97.9	97.9	93.1	97.9	93.1	98.6	99.3
PLP19$_\Delta$	98.6	98.6	97.2	98.6	97.2	97.9	98.6

Table 3.4. Performance of the A-AGR system for the voice source and vocal tract related features. Symbol Δ denotes use of both static and delta coefficients, e.g. for PLP19$_\Delta$ in total 19+19=38 coefficients were used. For LPCCs energy coefficient was not determined.

3.3.3 Vocal Tract Related Features

We studied performance of two types of the vocal tract related features: (a) formant features and (b) cepstral features under varying conditions, and the obtained results are presented in Table 3.4.

3.3.3.1 Formant Related Features

The results obtained under controlled (protocol **0**), degraded and adverse conditions (protocol **A**) showed that performance of formant related features, namely first four formant frequencies and bandwidths (F1-F4+B1-B4) significantly decreases with the severity of the conditions and is inferior to performance of F0 under all conditions. Further, the formant features were constantly worse than the cepstral features (which is consistent with the previous studies [69]), especially under noisy conditions. It is due possibly to unreliable estimate of the formant frequencies and bandwidths. The tuning of SVM parameters (γ, C) using development set specific for particular testing conditions (protocol **B**) slightly decreases performance of formant features. The addition of noisy data to the training set (protocol **C**) shows a trend similar as for F0.

3.3.3.2 Cepstral Features

We analyzed the cepstral features: LPCCs, MFCCs and PLPs, with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients. The comparison of the results obtained under controlled (protocol **0**), degraded and adverse conditions (protocol **A**) is provided in Figure 3.2. First, it can be observed that performance of all three cepstral features increases with the number of exploited cepstral coefficients under all conditions.[3] This trend is consistent with the

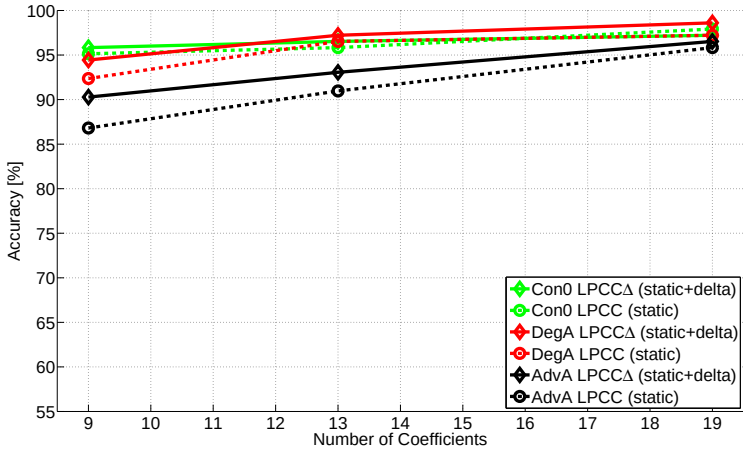
[3]The characteristics shown in Figure 3.2 are almost flat under controlled matched conditions (Con_0). This is a consequence of presenting results for file accuracy. However, the slight improvements in performance were observed for frame accuracy. For instance, the following frame accuracies: 81.2%, 84.6%, 86.4% and 87.0% were obtained for PLP9$_\Delta$, PLP13$_\Delta$, PLP19 and PLP19$_\Delta$, respectively.

results obtained on clean-condition data by Wu *et al.* [69]. However, it is important to note that the increase of number of cepstral coefficients aids in performances significantly more for degraded and adverse than controlled conditions. This leads to the conclusion that detailed modelling of spectrum is more crucial for noisy than clean-condition recordings. Second, the use of the delta coefficients in addition to the static coefficients further improved performance for all three cepstral features and under all conditions.[3] This observation is consistent with the results obtained on clean-condition data by Fussell *et al.* [18]. Consequently, the system employing 19 static and delta coefficients under mismatch noisy conditions can almost approach the performance as under clean matched conditions. Furthermore, LPCCs come out as more stable features than MFCCs and PLPs, in the sense that the amount of degradation in performance due to reduction of number of cepstral coefficients is significantly lower for LPCCs than for MFCCs and PLPs. This is possibly owing to the differences in characterizing a smooth spectral envelope by these features. In case of LPCCs, spectral peaks of the short-term spectrum are modelled directly, whereas in case of MFCCs and PLPs, the spectrum resulting from human auditory related processing is represented. Additionally, what can be useful for AGR, the estimation of the spectral peaks based on the linear prediction is affected by F0 for voiced speech segments, since formant frequencies and bandwidths are sensitive to the value of F0. This needs further investigation.

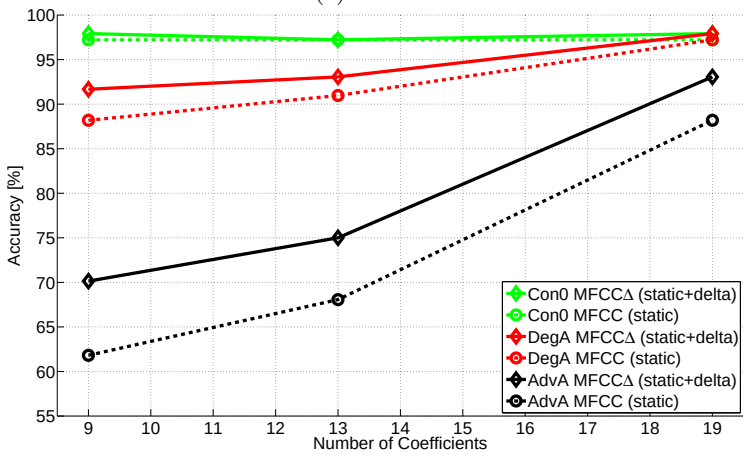
The comparison of the results obtained for protocol **B** (tuning of SVM parameters on noisy development data) with respect to the results obtained for protocol **A** (mismatched conditions) is provided in Figure 3.3. The tuning of SVM parameters (γ, C) using development set specific for the particular testing conditions slightly improves the performance of the cepstral features with lower number of coefficients.

The comparison of the results obtained for protocol **C** (multi-condition training) with respect to the results obtained for protocol **A** (mismatched conditions) is presented in Figure 3.4. Not surprisingly, the performance of the system under noisy conditions improved with multi-condition training. It can be observed that with multi-condition training: (a) performance of the system is less dependent upon the number of cepstral coefficients (i.e the amount of spectral details that are modelled); (b) the system employing only static coefficients yields performance closer to the system using both static and delta coefficients; and (c) using of both static and delta coefficients further improves performance of the system. Thereby, it can be hypothesized that modelling of more spectral detail and use of dynamic features is more important under mismatched conditions.

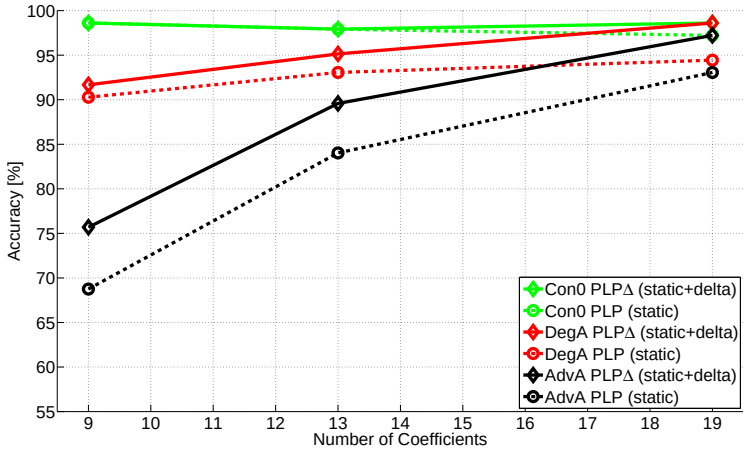
As a result, in the rest of the thesis, we report our studies for cepstral features with 19 coefficients including both static and delta coefficient to the feature vector. In such setup, as summarized in Table 3.4, all three cepstral representations provided similar performance under clean conditions, however PLPs were slightly better than MFCCs and LPCCs under noisy conditions. Moreover, under clean conditions the performance of F0 and cepstral features are comparable. However, under noisy conditions cepstral features yield more robust system. Also, multi-condition training helps more the cepstral-based system compared to F0-based system.



(a) LPCCs

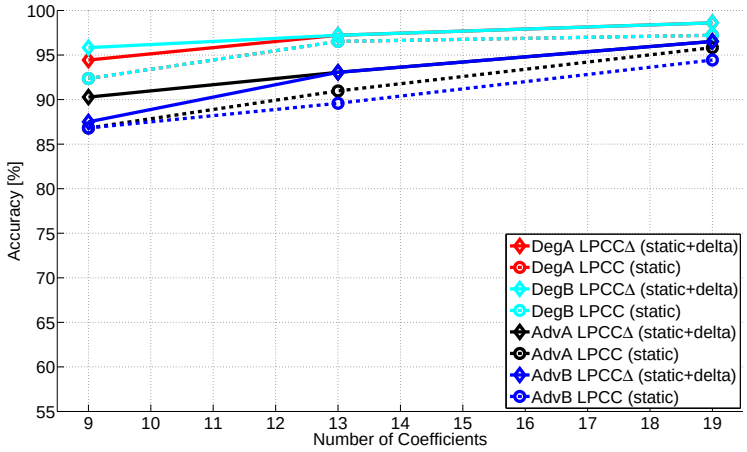


(b) MFCCs

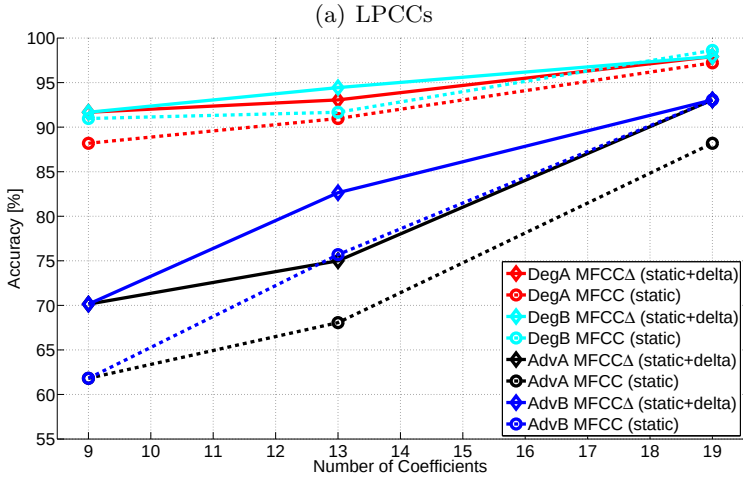


(c) PLPs

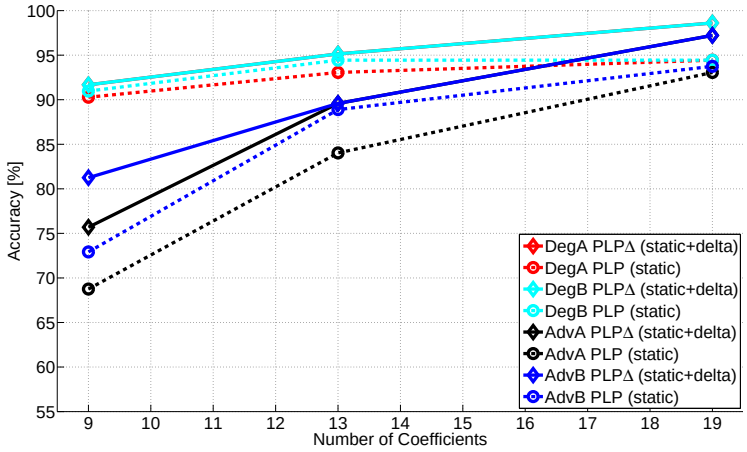
Figure 3.2. Performance of the A-AGR system for the cepstral features with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector under degraded (Deg_A) and adverse (Adv_A) conditions.



(a) LPCCs

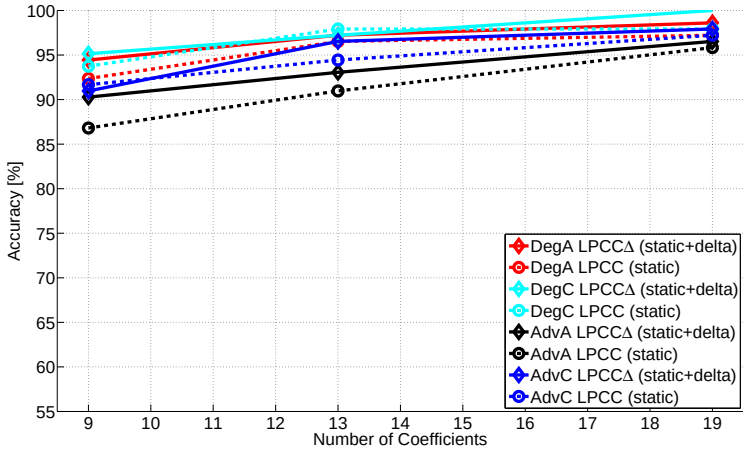


(b) MFCCs

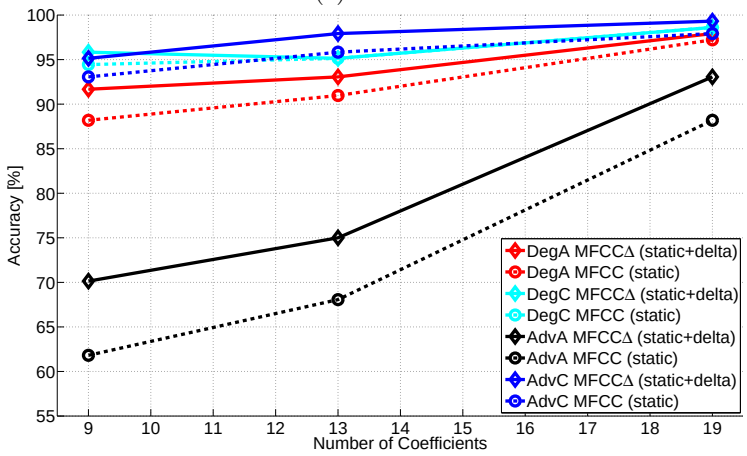


(c) PLPs

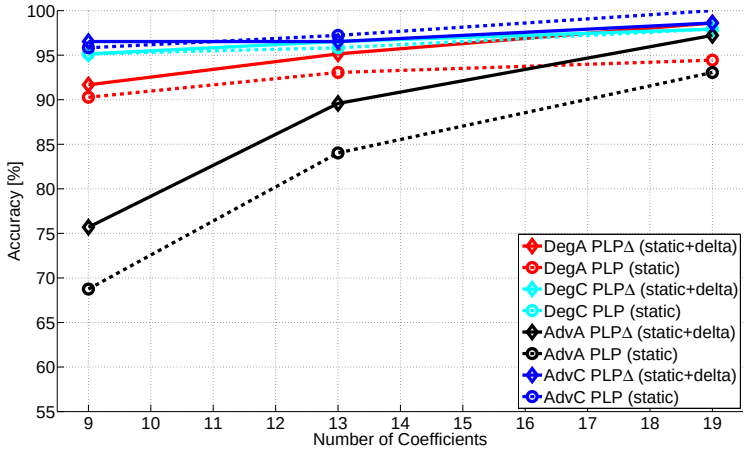
Figure 3.3. Comparison of performance of the A-AGR system across protocols **A** and **B**, i.e. training and development on controlled data vs. training on controlled data and development on noisy data.



(a) LPCCs



(b) MFCCs



(c) PLPs

Figure 3.4. Comparison of performance of the AGR system across protocols **A** and **C**, i.e. training and development on controlled data vs. training on controlled+noisy data (multi-condition training) and development on noisy data.

3.3.4 Audio Cue Integration

We analyzed integration of the voice source (F0) with the two types of vocal tract related features: (a) formant features and (b) cepstral features, and the obtained results are presented in Table 3.5. The low level integration of F0 with the cepstral features constantly yielded better system than integration of F0 with the formant features. This is consistent with the inferior results obtained for formant features in the previous section. Thus, further discussion is led for F0 intergated with the cepstral features.

In case of protocol **A** (mismatch conditions), high level integration yielded more resilient system than low level integration for all integrated features. While, for high level integration, comparing different methods of combining the classifiers outputs, sum and product rules provided higher recognition rates than max rule and, in addition, sum rule was superior to product rule. Results for integration of F0 and LPCCs with respect to employed combination rule (sum, product or max rule) are presented in Figure 3.5. For mismatch conditions (protocol **A**), integration of F0 with each of the three cepstral features provides identical high accuracy (99.3%) under degraded conditions, and combination of F0 with LPCC yields the best system (97.9%) under adverse conditions. In case of protocol **B**, when SVM classifier weights were determined based on development set specific for the test conditions, integration of F0 with LPCC yielded overall the best system with almost perfect recognition (99.3%) under both degraded and adverse conditions. In this case, the integration of audio features increased performance when compared to the best single feature system by 0.7% and 2.8% under degraded and adverse conditions, respectively. In order to compare the importance of voice source and vocal tract related features in the correct classification under varying conditions, we analyzed weights that were assigned to each type of features. The weights for F0 when integrating F0 with the cepstral features at high level using linear weighted summation are specified in Table 3.6. When the weights were determined based on the controlled condition development set (protocol **0** and **A**), higher importance was assigned to the voice source (weights 0.65-0.7) than for vocal tract related features (weights 0.3-0.35). This is not suprising, since F0 obtained perfect recognition under controlled conditions. However, for the development set specific for the test conditions (protocol **B**), higher importance was given to vocal tract related features under adverse conditions. In case of multi-conditional training, both types of features were almost equally important in the correct classification.

Finally, the studies on different training strategies revealed that instead of using both clean and noisy data to train classifier (protocol **C**), it may be a better strategy to use exclusively clean condition data for training, and then employ noisy data to estimate classifier weights (protocol **B**). In all studied cases, recognition rate was never reduced by determining weights on the development set specific for the test conditions. However, the multi-condition training (protocol **C**) decreases the performance for instance for the low level integration under degraded conditions and for LPCCs under adverse conditions.

Integrated Features		IL	Accuracy [%]						
			Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
			Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	F1-F4+ B1-B4	Low	99.3	96.5	93.1	96.5	92.4	95.1	95.1
F0	$LPCC18_\Delta$	Low	99.3	99.3	96.5	99.3	96.5	98.6	99.3
		High	100	99.3	97.9	**99.3**	**99.3**	99.3	97.9
F0	$MFCC19_\Delta$	Low	100	98.6	92.4	98.6	92.4	98.6	98.6
		High	99.3	99.3	93.1	99.3	93.1	99.3	97.9
F0	$PLP19_\Delta$	Low	99.3	99.3	94.4	99.3	94.4	98.6	98.6
		High	100	99.3	96.5	99.3	97.2	99.3	97.9

Table 3.5. Performance of the A-AGR system for F0 integrated with the vocal tract related features. Abbreviations: 'IL'=Integration Level.

Integrated Features		IL	Weight for F0 (w_{F0})						
			Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
			Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	$LPCC18_\Delta$	High	0.65	0.65	0.65	0.6	0.35	0.5	0.5
F0	$MFCC19_\Delta$	High	0.65	0.65	0.65	0.7	0.55	0.5	0.4
F0	$PLP19_\Delta$	High	0.7	0.7	0.7	0.65	0.25	0.5	0.4

Table 3.6. Wieghts obtained for F0 during integration of F0 with the cepstral features at high level using linear weighted summation. The corresponding wieght for the cepstral features is equal to $w_{CEP} = 1 - w_{F0}$. Abbreviations: 'IL'=Integration Level.

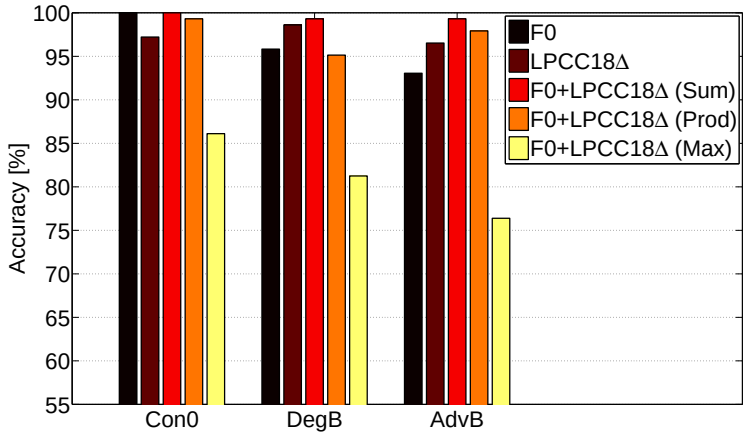


Figure 3.5. Comparision of performance of the A-AGR system for single and integrated the voice source (F0) and the vocal tract related features (LPCCs) under controlled (protocol **0**), degraded and adverse (protocol **B**) conditions. The high level integration is analyzed with respect to employed integration rule (sum, product, max).

3.4 Summary and Conclusions

In this chapter, we presented evaluation of different voice source and vocal tract related features for robust automatic gender recognition. Through studies performed on the BANCA corpus comprising datasets of varying complexity (controlled, degraded and adverse), we showed that:

Frame Selection

- Modelling only voiced speech frames improves the robustness of the AGR system towards mismatched conditions for both the voice source (F0) and vocal tract related features (LPCCs, MFCCs, PLPs);

Voice Source Related Features

- F0 can provide perfect recognition under controlled conditions, however the performance can degrade under noisy conditions;

Vocal Tract Related Features

- The cepstral features (LPCCs, MFCCs, PLPs) are superior to formant related features (formant frequencies and bandwidths);
- The performance of the AGR system is less sensitive to the number of cepstral coefficients (i.e. the amount of spectral details being modelled) under controlled conditions or with multi-condition training;
- Modelling of higher spectral details and the use of both static and dynamic features makes the system robust towards noisy conditions;

Voice Source vs. Vocal Tract Related Features. Audio Cue Integration

- F0 and the cepstral features provide similar performance under clean conditions, but cepstral features yield robust system in noisy conditions; PLPs were slightly better than MFCCs and LPCCs under noisy conditions.
- Integration of F0 with the cepstral features performed better than integration of F0 with the formant related features;
- When integrating audio cues, F0 was given more importance under clean conditions, however in noisy (adverse) conditions giving more importance to cepstral features yields a better system;
- The sum rule was found to be superior to product rule and max rule; the linear weighted summation of F0 and LPCCs yielded overall the best AGR system achieving almost perfect recognition (99.3%) under both degraded and adverse conditions when the system was exclusively trained on clean conditions data and the weights were adjusted on noisy data (protocol **B**).

Chapter 4

Vision-Based AGR Studies

In this chapter, we present experimental studies on different visual features for AGR task. The evaluation of low-dimensional representations of face images, such as eigenface and fisherface features, is performed in the framework of the V-AGR system described in Chapter 2. First, motivation of the studies is discussed and the main objectives are defined in Section 4.1. Then, experimenal setup is specified in Section 4.2, and results obtained under varying conditions are presented and discussed in Section 4.3. Finally, the main conclusions drawn from the studies are summarized in Section 4.4.

4.1 Motivation and Objectives

In early studies in vision-based gender recognition, different feature representations of face images have been proposed, such as characteristics extrated from raw image by ANNs or geometrical features, such as eyebrow thickness or nose width (see discussion in Section 2.3). Nevertheless, these features provided unsatisfactory performance and were computationally expensive. Latest studies in Automatic Face Recognition (AFR) suggested the low-dimensional representation of face images obtained by means of component analysis techniques as more robust and efficient. The features obtained by principal component analysis (eigenface features) and linear discriminant analysis (fisherface features) are the two most commonly used low-dimensional representations of face images for AFR. The thorough evaluation of these feature for AFR task showed that the eigenfaces provides high recognition rate when constant face pose and lightning are preserved and tends to underperform under varying conditions. Further, the fisherfaces achieved the superior performance compared to the eigenfaces in case of the large variation in lightning and facial expressions. The suitability of the eigenface features for the AGR problem was experimentally confirmed on clean condition data by Walawalkar *et al.* [68]. However, for automatic gender recognition the low-dimensional representations have been seldom evaluated under varying conditions. Alongside with the intention of studying audio-visual AGR, the above statement motivated us to evaluate performance of

the eigenfaces and fisherfaces under varying conditions (results are presented in Section 4.3). In this work, we address the following questions:

1. *What is the effect of varying conditions on the performance of the eigenface and fisherface features?* Which type of visual features is more robust towards varying conditions?

2. *What is the effect of training data conditions on the AGR accuracy?* Is a better strategy to train the AGR system on clean-condition or multi-condition (clean+noisy) data? How much the quality of data influences the gender information present in the visual features?

4.2 Experimental Setup

In this section, we first present the visual part of the BANCA database used to evaluate performance of the V-AGR system followed by information about data preprocessing and experimental setup.

4.2.1 Database

The V-AGR system was evaluated on a visual part of the BANCA database (English corpus) comprising datasets of varying complexity [4]. The visual data acquisition was performed using two cameras (poor-quality and good-quality) under three different types of conditions:

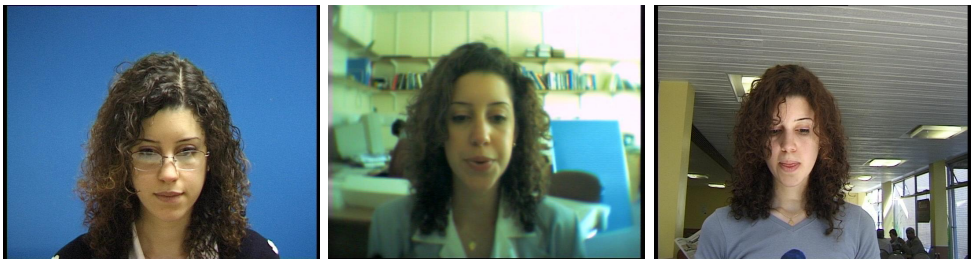
1. *Controlled*: good-quality camera, uniform background and stable lighting;

2. *Degraded*: poor-quality camera, non-uniform background;

3. *Adverse*: good-quality camera, arbitrary conditions [35].

Examples of images from the BANCA database collected under controlled, degraded and adverse conditions are presented in Figure 4.2, and a few more example images can be found in Appendix A.

In order to evaluate the performance of V-AGR system under varying conditions and determine the best strategy of training, we use the same four protocols defined earlier for A-AGR system study. First, the **0** protocol with matched controlled training and test conditions is established. Then, three additional protocols: **A**, **B**, **C** are defined which differ with respect to the quality of data used for training, development and testing. The protocol **A** with mismatched training and test data conditions, contains controlled condition data for training and development, and noisy condition data for testing. Next, the noisy data are used for development (protocol **B**) and, finally, are added to the training set (protocol **C**, multi-condition training). A summary of the protocols is given in Table 4.1. The BANCA database contains recordings collected from 52 subjects, and the same number of known and



(a) Controlled conditions (b) Degraded conditions (c) Adverse conditions

Figure 4.1. Examples of images from the BANCA database [4] collected under controlled, degraded and adverse conditions. More images are presented in Appendix A.

(a) Controlled conditions (b) Degraded conditions (c) Adverse conditions

Figure 4.2. Automatically detected and extracted face regions from the BANCA images presented in Figure 4.2.1.

unknown subjects is used for testing performance of the V-AGR system. More information about the division of subjects and details about experimental setup can be found in Appendix A.

4.2.2 Analysis of Visual Data

In order to extract a face region from an image, first an automatic frontal face detector performing geometric normalization of the image in order to align eyes was applied. Then, each image was cropped to a size of 64x80 in order to preserve only deteced face region. Examples of automatically detected and extracted face regions from the BANCA images are presented in Figure 4.2.1. For this purpose, tools provided in the Torch3vision library were used [1]. The obtained face images were first projected into an eigenspace and, then, fisherspace. The *Principle Component Analysis* (PCA) was performed to obtain eigenface features. The number of features was chosen to capture 99% of the data variations which, in this case, corresponds to the first 116 eigenvectors. When the number of eigenvectors was increased above this value no significant change in performance was observed on the development data set. Then, the *Linear Discriminant Analysis* (LDA) was conducted to determine fisherface features. Due to the fact that the informative part of the LDA features is encoded in the first $n-1$ vectors, where n is the number of classes, for gender recognition problem each image was represented using only one feature.

Protocol	Set ID	Conditions		
		TRAIN	DEV	TEST
0	Con_0	**Controlled**	**Controlled**	**Controlled**
A	Deg_A	**Controlled**	**Controlled**	Degraded
	Adv_A	**Controlled**	**Controlled**	Adverse
B	Deg_B	**Controlled**	Degraded	Degraded
	Adv_B	**Controlled**	Adverse	Adverse
C	Deg_C	**Controlled**+Degraded	Degraded	Degraded
	Adv_C	**Controlled**+Adverse	Adverse	Adverse

Protocol	Item	TRAIN	DEV	TEST
0,**A**,**B**,**C**	Subjects Σ(F,M)	26(13,13)	16(8,8)	36(18,18)
	# Images per File	5	5	5
0,**A**,**B**	# Files	104	64	144
C	# Files	104+64	64	144

Table 4.1. Experimental setup for protocols **0**, **A**, **B** and **C**. Abbreviations and symbols: 'Σ'=Total, 'F'=Females, 'M'=Males. Details can be found in Appendix A.

4.2.3 Classification

As in case of A-AGR system, we employed the SVMs implemented in the LIBSVM library to perform gender classification [9] (details in Section 2.4). The RBF kernel was used in case of multi-dimensional feature vectors, and the linear kernel in case of one-dimensional feature vectors. The parameters of the SVMs (error penalty C) and the RBF kernel (variance γ) were estimated on the development set. However, as mentioned earlier, in case of simple almost linearly separable problems, the choice of C is of small importance. The a posteriori class probabilities are estimated using the Platt's method [47].

4.2.4 Performance Evaluation

The V-AGR system is evaluated on the test set and its performance is expressed as percentage of correct classification, i.e. *classification accuracy*. Each image is assumed as a separate sample and is classified independently. As a measure of confidence with which a sample is assigned to a particular class, the a posteriori class probability is chosen. The results are reported at a file level based on 5 images extracted from each video file ($K_V = 5$). The decision about gender for a single file is obtained by summing[1] the image values of the a posteriori probabilities for each class for 5 images, and then choosing the class with the higher score.

[1]Preliminary experiments performed on the development set showed that summation outperforms product and max function.

Feature	Accuracy [%]						
	Protocol **0**	Protocol **A**		Protocol **B**		Protocol **C**	
	Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
Eigenfaces	93.1	82.6	73.6	86.8	81.9	90.3	91.0
Fisherfaces	95.1	82.6	81.3	82.6[1]	81.3[1]	90.3	90.3

Table 4.2. Performance of the V-AGR system for the eigenface and fisherface features. [1]The results for the protocol **B** are identical as for **A**, since the same values of parameter C for the SVMs with the linear kernel were determined.

4.3 Results and Discussion

In case of the V-AGR system, we studied two low-dimensional representations of face images under varying conditions, namely the eigenfaces and fisherfaces. The results obtained during the experiments are presented in Table 4.2. When the AGR system was trained exclusively on clean conditions data (protocol **0** and **A**), both features obtained good recognition rates under controlled conditions, however their performances were heavily affected by degradation of conditions. The fisherfaces provided superior performance to the eigenfaces, especially under adverse conditions. This is consistent with the results obtained by Belhumeur *et al.* for AFR problem [3]. It is due mostly to the fact that the fisherfaces are more robust to large variation in lightning and facial expressions than the eigenfaces.

The tuning of SVM parameters (γ, C) using development set specific for particular testing conditions (protocol **B**) significantly improved performance of the eigenfaces, such that they achieved better performance than the fisherfaces under degraded and adverse conditions. On the other hand, no effect on performance of the fisherfaces was observed. It is due to the fact that the fisherfaces are one-dimensional feature for AGR and the SVMs with the linear kernel are used. In consequence, influence of using noisy condition development set is limited to adjusting the error penalty C. In our studies, the same value of the parameter C was determined for protocol **A** and **B**.

The multi-condition training (protocol **C**) significantly improved performance of both the eigenfaces and fisherfaces, since new example of images with conditions (such as lighting, background etc.) specific for test set were introduced to the system during training. The eigenfaces and fisherfaces obtained identical performance under degraded conditions, and the eigenfaces slightly outperformed the fisherfaces under adverse conditions. It is important to note that under degraded and adverse conditions for protocol **C**, the eigenfaces and fisherfaces achieved performance close to results obtained under controlled conditions (**0**). From this observation, it can be concluded that matching of training and test conditions is important for both low-dimensional representations.

4.4 Summary and Conclusions

In this chapter, we presented evaluation of two low-dimensional representations of face images, namely the eigenfaces and fisherfaces for robust automatic gender recognition. The studies performed on the BANCA database comprising datasets of varying complexity (controlled, degraded and adverse) showed that:

- Performance of the AGR system significantly decreases for both the visual features, i.e. eigenfaces and fisherfaces, with degradation of the conditions. Furthermore, similar to earlier AFR studies, the fisherface were found to be more robust to unseen and varying conditions;
- The use of noisy development set to tune SVM parameters aids in performance for the eigenfaces; however, no effect on performance of the fisherfaces was observed due to the fact that this feature has only one dimension for AGR problem and the SVMs with linear kernel were used;
- The multi-condition training significantly improved performance for both the eigenfaces and the fisherfaces with the eigenfaces being slightly better under adverse conditions. In short, the match between training and test data conditions is important for both the visual features.

Chapter 5

Audio-Visual AGR Studies

This chapter presents experimental studies on integration of audio and visual features for AGR task. The evaluation of the different integration methods of audio and visual cues is performed in the framework of AV-AGR system described in Chapter 2. First, motivation of the studies is discussed and the main objectives are defined in Section 5.1. Then, performance for the selected audio and visual features is compared in Section 5.2. The experimental setup is specified in Section 5.3 and the obtained results are presented and discussed in Section 5.4. Finally, the major conclusions drawn from the studies are summarized in Section 5.5.

5.1 Motivation and Objectives

The previous studies on the AGR problem have considered the use of single modality: audio or vision (see discussion in Section 1.1). Moreover, these works have been performed mainly under clean conditions and the robustness of AGR systems in real-world scenarios was seldom considered. However, in most typical applications, both audio and visual signals are available. Therefore, ideally, an AGR system should be able to exploit both modalities to improve robustness of the system. Since each modality has different characteristics, audio-visual cues can provide a more comprehensive description of a subject than a single modality. Moreover, integration of the cues may yield a AGR system that is resilient to the degradation of both, or even to temporal unavailability of one of the input signals. This motivated us to conduct the audio-visual AGR studies and evaluate performance of the AV-AGR system under varying conditions. Since, the choice of the most suitable audio and visual features was thoroughly discussed in the framework of the uni-modal AGR systems in the last two chapters, here we focus on different methods of fusing the audio and visual cues at the classifier level with emphasis on the following practical questions:

1. *What is the effect of fusing the audio and visual information on the AGR accuracy?* How much complementary information is provided by the audio and visual cues?

2. *What is the best method of integrating the audio and visual cues?* What is the effect of using different algebraic combination methods, such as sum, product or max rule on the AGR accuracy? What is the effect of using different weighting schemes, such as equal and unequal weighting on the AGR accuracy?

3. *What is the effect of varying conditions on the performance of the AV-AGR system?* Combination of which audio and visual feature is more robust towards varying conditions?

4. *Which type of information, audio or visual, is more important in the correct classification under varying conditions?*

5. *What is the effect of training data conditions on the AV-AGR accuracy?* Is a better strategy to train the AGR system on clean-condition or multi-condition (clean+noisy) data?

5.2 Comparision of Audio and Visual Features

In this section, we first compare the results obtained in the framework of the uni-modal AGR systems in order to provide the background for further discussion about integration of audio and visual cues. Performance of the A-AGR and V-AGR system for selected features is summarized in Table 5.1. During the cue integration studies, we evaluate different combinations of both types of the audio features, namely the voice source (F0) and vocal tract related features (PLPs), with the two visual features (eigenfaces, fisherfaces). The choice of PLPs as a representative of the vocal tract related features is motivated by the fact that PLPs provided better performance under noisy conditions than MFCCs and LPCCs. Further, the best setup for PLPs is used, i.e. a feature vector contains 19th static and delta coefficients (see Section 3.3.3). When comparing results for the A-AGR and V-AGR system trained exclusively on controlled data (protocol **0** and **A**), it can be observed that the audio features are superior to the visual features under all conditions. Moreover, performance of all features decreased with degradation of data quality, and the drop was much greater for the visual than the audio features. Under adverse conditions, the difference between the PLPs and fisherfaces is equal to 15.9%, and between the PLPs and eigenfaces to 23.6%.The tuning of SVM parameters using the development set specific for particular testing conditions (protocol **B**) aided in performance only in case of the eigenfaces. Further, the addition of noisy data to the training set (protocol **C**), decreased performance for the audio features under degraded conditions, and increased in all other cases. The improvement is most significant for the visual features. The difference in performance between the PLPs and fisherfaces is reduced to 8.3%, and between the PLPs and eigenfaces to 7.6% (see Table 5.1).

Summarizing, the visual feature are much less robust to varying conditions than audio features. While comparing results at file level, not insignificant is the amount

Feature	Accuracy [%]						
	Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
	Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	100	95.8	93.1	95.8	93.1	95.1	94.4
$PLP19_\Delta$	98.6	98.6	97.2	98.6	97.2	97.9	98.6
Eigenfaces	93.1	82.6	73.6	86.8	81.9	90.3	91.0
Fisherfaces	95.1	82.6	81.3	82.6	81.3	90.3	90.3

Table 5.1. Performance of the A-AGR and V-AGR system for selected features.

of data samples available for each modality. The audio modality provides large amount of data ($K_A = 129$), whereas visual modality offers smaller number of relatively richer samples ($K_V = 5$). While comparing results obtained at sample level (for a single frame or image), the difference in performance between the audio and visual features is smaller[1] than at file level, however the A-AGR system consistently outperformed the V-AGR system under all conditions.

5.3 Experimental Setup

In this section, we provide information about the database and the AV-AGR system setup used during the experiments. In case of the analysis of audio and visual data the same setup was preserved as described in Sections 3.2.2 and 4.2.2, respectively.

5.3.1 Database

The AV-AGR system was evaluated simultaneously on both the audio and the visual part of the BANCA database (English corpus) [4]. As metioned in the previous chapters, data acquisition was performed using two cameras and two microphones under three different types of conditions:

1. *Controlled*: good-quality microphone and camera, clean audio conditions, uniform background and stable lighting;

2. *Degraded*: poor-quality microphone and camera, stable audio conditions, non-uniform background;

3. *Adverse*: good-quality microphone and camera, background noise, arbitrary conditions [35].

[1]Difference in accuracy at sample level (frame, image) between the best audio and the best visual features is equal to 1.1%, 8.7% and 6.2% under controlled, degraded and adverse conditions (protocol **0** and **A**), respectively.

Examples of images from the BANCA database collected under controlled, degraded and adverse conditions are presented in Figures 4.2.1. In order to evaluate performance of AV-AGR system under varying conditions and determine the best strategy of training, the four identical protocols are used as in case of the A-AGR and V-AGR system. A summary of these protocols is given in Tables 3.1 and 4.1 for the audio and the visual part of the database, respectively. More information about division of subjects and details about experimental setup can be also found in Appendix A.

We report the results at a file level based on 1.3s of voiced speech segment and 5 images extracted from each video file. The decision about gender for a single file is obtained based on the a posteriori class probabilities estimated at file level for each modality as described in Section 2.5.

5.3.2 System Setup

As described in Section 2.5.1, the AV-AGR system in this work is created by fusing evidences from the two modalities at the *high level*, after the single-cue classification based on SVMs has been performed. As a measure of confidence with which a subject was assigned to a particular class, the a posteriori class probability was chosen. The confidence measures with which a subject was assigned to a particular class from the A-AGR and V-AGR system were integrated using a weighted combination method, such as sum, product or max rule to provide the final decision based on both modalities (details in Section 2.5.1). Additionally, we considered equal or unequal weighting of modalities during the experiments, and the latter were performed in order to answer the question of different importance of the modalities in the correct classification. In case of unequal weighting of modalities, the weights were determined to maximize performance of the AV-AGR system on the development set. Each weight can take a value between zero and one with the step of 0.05. In other cases, the same setup was preserved as described in Sections 3.2.3 and 4.2.3, respectively.

5.3.3 Performance Evaluation

The AV-AGR system is evaluated on the test set and its performance is expressed as percentage of correct classification, i.e. *classification accuracy*. As in case of unimodal AGR systems, each frame of audio signal and each image is assumed as a separate sample and is classified independently. As a measure of confidence with which a sample is assigned to a particular class, the a posteriori class probability is chosen. We report the results at a file level based on: (a) 1.3s of speech segment or voiced speech segment ($K_A = 129$), and (b) 5 images ($K_V = 5$), extracted from each video file. The decision about gender for a single file is obtained according to Equation 2.41 where the evidence d_{ij} estimated for the i-th classifier and the j-th class based on a single sample is substituted by the evidence d_{ij}^{ALL} determined based on all samples in audio and visual sequences according to Equation 2.45 for $j \in \{Female, Male\}$ and $i \in \{Audio, Vision\}$ as described in Section 2.5.1.

5.4 Results and Discussion

For the AV-AGR system, we evaluated different combinations of audio (F0, PLPs) and visual (eigenfaces, fisherfaces) features. In the framework of protocol **A** (mismatch conditions), we compared different algebraic methods of combining estimates of the a posteriori probabilities provided by the uni-modal AGR systems, such as sum, product or max rule. The sum and product rule were vastly superior to the max rule for all combinations of audio and visual features and under all conditions. In addition, the sum rule provided better performance than the product rule for F0 integrated with each of the visual features, and comparable performance as product rule for PLPs integrated with the visual features. Figure 5.1 shows the comparison of the results obtained for combinations of F0 and $PLP19_{\Delta}$ with the eigenfaces under varying conditions.

Moreover, during the experiments, we considered equal or unequal weighting of modalities. Not surprisingly, the unequal weighting outperforms the equal weighting of modalities for all combinations of features and under all conditions. The difference in the performance between these two types of weighting increases with the severity of conditions. Figure 5.2 shows the comparison of the performance for equal and unequal weighting using sum rule. In the rest of the discussion, we restrict ourselves to sum rule with unequal weights determined on the development set.

The performance of the AV-AGR system for different types of features under varying conditions is specified in Table 5.2 and compared with performance of the uni-modal AGR systems in Figure 5.3. The combinations of F0 with each of the visual features attained perfect recognition rate under controlled conditions, however under noisy conditions their performances are inferior to results obtained for integration of PLPs with the visual features. This characterstic is clearly inherited from the A-AGR system based on the voice source related features (F0). The inte-

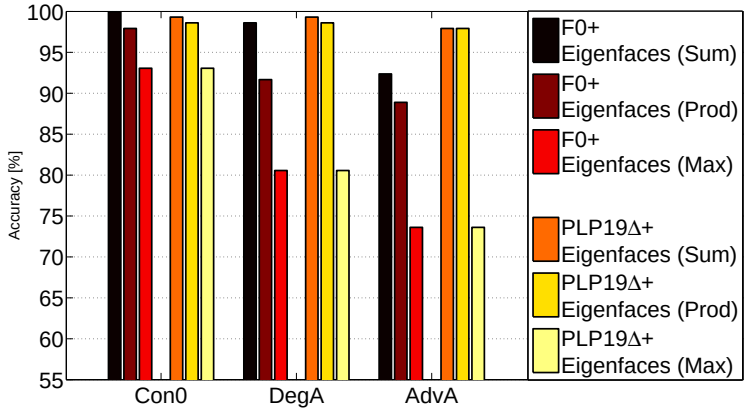


Figure 5.1. Performance of the AV-AGR system with respect to employed integration rule (sum, product, max) under controlled, degraded and adverse conditions (protocols **0** and **A**). The presented results were obtained for unequal weighting of modalities.

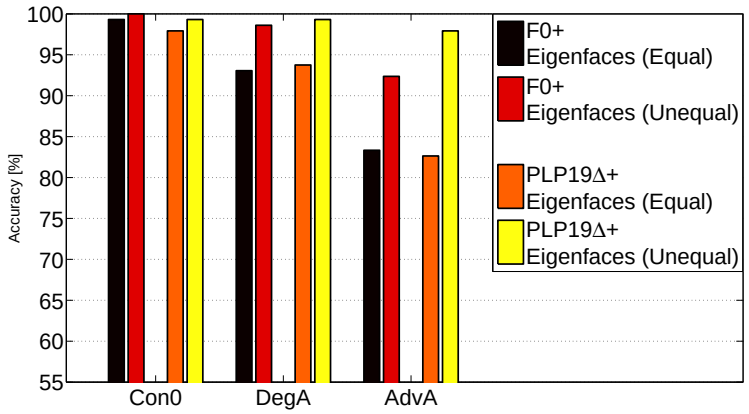


Figure 5.2. Performance of the AV-AGR system with respect to employed type of modality weighting (equal or unequal weighting) under controlled, degraded and adverse conditions (protocols **0** and **A**). The presented results were obtained for the sum rule.

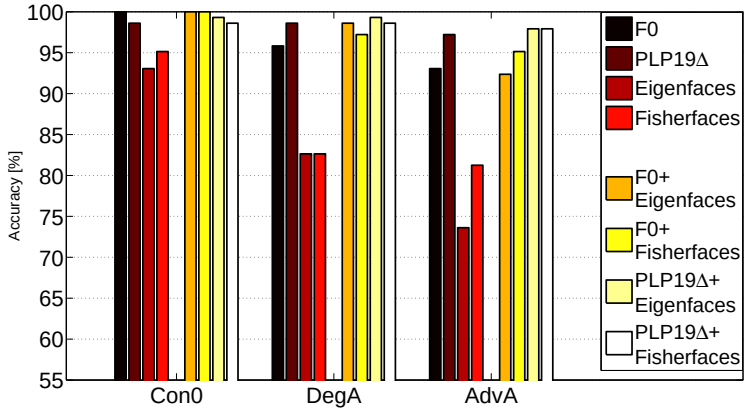


Figure 5.3. Comparision of performance of the uni-modal AGR systems and the AV-AGR system under controlled, degraded and adverse conditions (protocols **0** and **A**). The presented results were obtained for the audio and visual features integrated using the linear weighted summation of modalities where the weights were determined based on the development sets.

gration of PLPs with each of the visual features yielded a resilient system that at least preserved the performance of the best uni-modal AGR system under all conditions. Integration of PLPs with the eigenfaces improved performance by 0.7% when comparing to the A-AGR system under both degraded and adverse conditions, and by 16.7% and 24.3% while comparing to the V-AGR system under degraded and adverse conditions, respectively. The relatively small improvement of performance compared to the A-AGR system and the large improvement of performance compared to the V-AGR system is due to the fact that the A-AGR system is superior to the V-AGR system under all conditions. As presented in Table 5.3, the audio features get higher weights than the visual features and their importance in correct

classification increases with degradation of the data. Future work will address the problem of improving the robustness of the V-AGR system.

When the weights were determined based on the development set specific for the test conditions (protocol **B**), importance of the audio features increased for all combination of the audio and visual features and under all conditions. For integration of the audio features with the eigenfaces, the visual information was entirely neglected under adverse conditions. The increase of the weights for audio features slightly improved results only for the combination of F0 with the eigenfaces, namely to the level of performance of the A-AGR system employing F0. However, at the same time, the performances of all other combinations of features decreased or, in the best case, remained unchanged. It may be disadvantageous because of the visual system.

Finally, when using both clean and noisy data to train the classifiers (protocol **C**), the performance of the AV-AGR system decreased under degraded conditions compared to protocol **A**, except for the integration of PLPs with the fisherfaces for which perfect recognition was attained and performance increased by 2.1% and 9.7% compared to the A-AGR and V-AGR system, respectively. On the other hand, the performance under adverse conditions increases for all combination of features compared to protocol **A**. The almost perfect recognition was obtained for the combination of PLPs with the fisherfaces and, in this case, performance of the AV-AGR system increased by 0.7% and 9.0% compared to the A-AGR and V-AGR system, respectively. The weights for audio features decreased for all combination of features and under all conditions compared to protocol **B**.

5.5 Summary and Conclusions

In this chapter, we studied different methods of integration the audio (such as F0 and PLPs) and visual (such as eigenfaces and fisherfaces) cues for robust automatic gender recognition. Recognition studies performed on the BANCA corpus comprising datasets of varying complexity (controlled, degraded and adverse) showed that:

- Combination of PLPs with the visual features yields a resilient system that at least preserved the performance of the best uni-modal AGR system under both clean and noisy conditions;

- Among the combination methods, the sum and product rules provide comparable results and both rules were vastly superior to the max rule; the unequal weighting of modalities yields better system than the equal weighting of modalities;

- The combinations of F0 with each of the visual features attained perfect recognition under controlled conditions, however the combination of PLPs with the visual features provided superior performance under noisy conditions;

Integrated Features		Accuracy [%]						
		Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
		Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	Eigenfaces	100	98.6	92.4	97.2	93.1	93.1	96.5
F0	Fisherfaces	100	97.2	95.1	97.2	93.8	97.9	97.9
$PLP19_\Delta$	Eigenfaces	99.3	99.3	97.9	98.6	97.2	97.2	98.6
$PLP19_\Delta$	Fisherfaces	98.6	98.6	97.9	98.6	97.9	100	99.3

Table 5.2. Performance of the AV-AGR system for protocols **0**, **A**, **B**, **C**. The presented results were obtained for the audio and visual features integrated using the linear weighted summation of modalities where the weights were determined based on the development sets. The corresponding weights for the audio features are presented in Table 5.3.

Integrated Features		Weight for audio features (w_{Audio})						
		Prot. **0**	Protocol **A**		Protocol **B**		Protocol **C**	
		Con_0	Deg_A	Adv_A	Deg_B	Adv_B	Deg_C	Adv_C
F0	Eigenfaces	0.80	0.80	0.80	0.90	1.00	0.50	0.90
F0	Fisherfaces	0.90	0.90	0.90	0.90	0.95	0.80	0.80
$PLP19_\Delta$	Eigenfaces	0.75	0.75	0.75	0.95	1.00	0.50	0.85
$PLP19_\Delta$	Fisherfaces	0.90	0.90	0.90	0.95	0.95	0.75	0.75

Table 5.3. Weights obtained for the audio features during integration of the audio (F0, PLPs) and visual features (eigenfaces, fisherfaces) using the linear unequal weighted summation. The corresponding weight for the visual feature is equal to $w_{Visual} = 1 - w_{Audio}$. The corresponding results for the AV-AGR system are presented in Table 5.2.

- For realistic scenarios, in case of availability a set of data of quality specific for testing conditions during a training phase, a better strategy is to perform a multi-condition training (protocol **C**).

Chapter 6

Conclusions

In this work, we studied a multi-modal AGR system based on audio and visual cues and studied its performance in realistic scenarios. First, in the framework of two uni-modal AGR systems, we analyzed robustness of different audio (pitch frequency, formant and cepstral representations) and visual (eigenfaces, fisherfaces) features under varying conditions. Then, we built an integrated audio-visual system by fusing information from each modality at the classifier level using different combination rules and type of weighting. Our studies were conducted on the BANCA database comprising datasets of varying complexity (controlled, degraded and adverse). In the framework of the uni-modal AGR systems, we showed that:

- the audio-based system is more robust than the vision-based system, and its resilience to noisy conditions is increased by modelling only the voiced speech frames;
- in case of audio, the cepstral features are superior to the pitch frequency and formant features, although the pitch frequency obtained perfect recognition under clean conditions; PLPs yields slightly better system compared to other cepstral features;
- for the cepstral features, modelling of higher spectral details and the use of both static and delta coefficients made the system robust towards noisy conditions;
- in case of vision, the fisherfaces outperformed the eigenfaces under degraded and adverse conditions.

When evaluating different cue integration methods for the audio-visual AGR systems, we showed that:

- the sum and product combination rule provide comparable results, and both rules are vastly superior to the max rule;
- the unequal weighting of modalities yields better system than the equal weighting of modalities.

Finally, the integration of audio and visual cues yielded a robust system that preserved the performance of the best modality in clean conditions and helped in improving performance in noisy conditions (combination of the PLPs and eigenfaces provided the very high performance of 99.3% and 97.9% under degraded and adverse conditions, respectively). Summarizing, the careful selection of the audio and visual features and integration of the multi-modal cues yielded the resilient AGR system applicable in practical applications.

6.1 Future Work

The work presented in this thesis can be extended in a number of directions:

- the audio-based AGR system:

  - in case of the formant related features, the formant frequency and bandwidth were studied. The addition of the third parameter, namely amplitudes of the formants may improve results obtained for the formant features;

- the vision-based AGR system:

  - performance of the system may be improved by usage of a more robust automatic face detector that will be invariant to pose and rotation of a face;

  - other possible representations of face regions may be investigated, for instance features based on a texture analysis such as the local binary patterns (LBPs) [44] which demonstrated the high performance in Automatic Face Recognition systems [35];

- the audio-visual AGR system:

  - other types of weights may be investigated, such as the dynamic weights which are determined separately for each sample such as inverse of entropy, or can be found using the Dempster-Shafer method [31].

Appendix A

BANCA Database

In this work, we used the BANCA database (English corpus) which contains datasets of varying complexity [4, 16]. The data acquisition was performed using two cameras and two microphones under three different types of conditions:

1. *Controlled*: good-quality microphone and camera, clean audio conditions, uniform background and stable lighting;

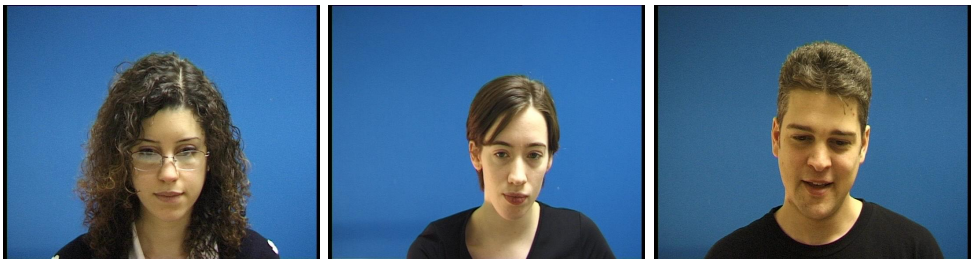
2. *Degraded*: poor-quality microphone and camera, stable audio conditions, non-uniform background;

3. *Adverse*: good-quality microphone and camera, background noise, arbitrary conditions [35].

Examples of images from the BANCA database collected under controlled, degraded and adverse conditions are presented in Figure A.1.

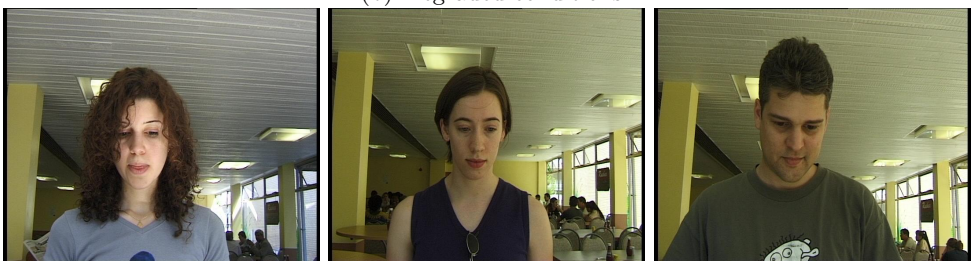
In each conditions, 4 sessions were scheduled during which 2 recordings from 52 subjects (26 females, 26 males) were collected. Subjects were asked to say a random 12 digit number, their name, address and date of birth [16]. The BANCA recordings were divided into the four random subsets: S_1, S_2, L_1, L_2, with respect to subject's id as it is presented in Table A.1. In order to evaluate the system on the same number of known and unknown subjects, only a half of the subjects, 26 (subsets S_1+L_1), were used for training. Then, all 52 subjects were split into two groups consisting of 16 (S_1+S_2) and 36 (L_1+L_2) subjects which were used for development and testing.

In order to evaluate performance of the system under clean conditions the **0** protocol (matched controlled training and test conditions) was established. Then, to determine a strategy of training that will yield the most robust system under varying conditions, three additional protocols: **A**, **B**, **C** were defined each in two versions: *Deg* and *Adv* for degraded and adverse conditions, respectively. As specified in Table A.2, the three protocols differ with respect to the quality of data used for training, development and testing. The idea was to first use the clean conditions data for training and development, and test the system under noisy conditions



(a) Controlled conditions.

(b) Degraded conditions



(c) Adverse conditions

Figure A.1. Examples of images from the BANCA database [4] collected under controlled, degraded and adverse conditions.

Subjects' Group ID	Number of subjects			BANCA Subjects' id	
	Total	Female	Male	Female	Male
S_1	8	4	4	03,06,16,22	32,34,35,40
S_2	8	4	4	07,11,14,17	30,33,45,51
L_1	18	9	9	01,05,08,15,19,21,23,25,26	28,29,38,41,42,46,47,48,52
L_2	18	9	9	02,04,09,10,12,13,18,20,24	27,31,36,37,39,43,44,49,50

Table A.1. Division of the BANCA database into the four random subsets with respect to subject's id ($id_{BANCA} = 10xx$, where xx is specified in the tabel) [4].

(protocol **A**, mismatched training and test conditions). In protocol **B**, the training is done with clean condition data and the parameters of the system are tuned with noisy development data. Finally, in protocol **C** both clean and noisy data are used for training (multi-condition training).

Protocol	Set ID	Conditions			BANCA Session			Subjects' Group		
		TRAIN	DEV	TEST	TRAIN	DEV	TEST	TRAIN	DEV	TEST
0	Con_0	**Con.**	**Con.**	**Con.**	1,2	3,4	3,4	S_1,L_1	S_1,S_2	L_1,L_2
A	Deg_A	**Con.**	**Con.**	Deg.	1,2	3,4	7,8	S_1,L_1	S_1,S_2	L_1,L_2
	Adv_A	**Con.**	**Con.**	Adv.	1,2	3,4	11,12	S_1,L_1	S_1,S_2	L_1,L_2
B	Deg_B	**Con.**	Deg.	Deg.	1,2	7,8	7,8	S_1,L_1	S_1,S_2	L_1,L_2
	Adv_B	**Con.**	Adv.	Adv.	1,2	11,12	11,12	S_1,L_1	S_1,S_2	L_1,L_2
C	Deg_C	**Con.**+Deg.	Deg.	Deg.	1,2+5,6	7,8	7,8	S_1,L_1+S_1,S_2	S_1,S_2	L_1,L_2
	Adv_C	**Con.**+Adv.	Adv.	Adv.	1,2+9,10	11,12	11,12	S_1,L_1+S_1,S_2	S_1,S_2	L_1,L_2

Protocol	Item	TRAIN	DEV	TEST
0,**A**,**B**,**C**	Subjects Σ(F,M)	26(13,13)	16(8,8)	36(18,18)
	Audio Data per File	1.5s	1.3s	1.3s
	# Images per File	5	5	5
0,**A**,**B**	# Files	104	64	144
	Audio Data Total	156s	84s	188s
	# Images Total	520	320	720
C	# Files	104+64	64	144
	Data Total	252s	84s	188s
	# Images Total	840	320	720

Table A.2. Experimental setup for protocols **0**, **A**, **B** and **C**. Abbreviations and symbols: 'Con.'=Controlled, 'Deg.'=Degraded, 'Adv.'=Adverse, 'Σ'=Total, 'F'=Females, 'M'=Males.

Bibliography

[1] Machine Vision Library - Torch3vision. Retrieved April, 2008 from <http://torch3vision.idiap.ch>.

[2] Manual for the audio processing toolkit - Snack v2.2.9. Retrieved April, 2008 from <http://www.speech.kth.se/snack/>.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Poree, E. Bailly-Bailliere, G. Matas, and B. Ruiz. Experimental Protocol on the BANCA Database. Technical report, 2002.

[5] Ch. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[6] V. Bruce, A. M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. Sex discrimination: how do we tell the difference between male and female faces? *Perception*, 22(2):131–152, 1993.

[7] R. Brunelli and T. Poggio. HyberBF networks for gender classification. In *Proceedings of the Image Understanding Workshop*, pages 311–314, San Mateo, CA, USA, January 1992. Morgan Kaufmann Publishers Incorporated.

[8] R. Campbell, S. B. Benson, P. J.and Wallace, S. Doesbergh, and M. Coleman. More about brows: How poses that change brow position affect perceptions of gender. *Perception*, 28(4):489–504, 1999.

[9] C-C. Chang and C-J. Lin. LIBSVM. Retrieved April, 2008 from <http://www.csie.ntu.edu.tw/cjlin/libsvm>.

[10] D. G. Childers and K. Wu. Gender recognition from speech. Part II: Fine analysis. *Journal of the Acoustical Society of America*, 90:1841–1856, October 1991.

[11] E. P. Chronicle, M. Y. Chan, C. Hawkings, K. Mason, K. Smethurst, K. Stallybrass, K. Westrope, and K Wright. You can tell by the nose – judging sex from an isolated facial feature. *Perception*, 24(8):969–973, 1996.

[12] R. O. Coleman. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech, Language, and Hearing Research*, 19:168–180, 1976.

[13] G. W. Cottrell and J. Metcalfe. Empath: face, emotion, and gender recognition using holons. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 564–571, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Incorporated.

[14] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience, 2nd edition, November 2000.

[16] E. Bailly-Baillire et al. The BANCA database and evaluation protocol. *LNCS*, 2688:1057–1072, 2003.

[17] G. Fant. *Speech sounds and features.* MIT Press, Cambridge, 1973.

[18] J. W. Fussell. Automatic sex identification from short segments of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 409–412, 1991.

[19] B. Gold and N. Morgan. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music.* John Wiley & Sons, 1999.

[20] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 572–577, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Incorporated.

[21] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Proceedings of the Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998.

[22] S. Haykin. *Neural networks: a comprehensive foundation.* Prentice–Hall, 2nd edition, 1998.

[23] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[24] Geoffrey E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1-3):185–234, 1989.

[25] M. Hirano, S. Kurita, and T. Nakashima. Growth, development, and aging of human vocal folds. *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, pages 23–43, 1983.

[26] J. Holmes and W. Holmes. *Speech Synthesis and Recognition.* Taylor & Francis Group, 2nd edition, 2001.

[27] Xuedong Huang and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[28] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

[29] L. Jie, B. Caputo, A. Zweig, J-H. Bach, and J. Anemuller. Object category detection using audio-visual cues. In *Proceedings of the International Conference on Computer Vision Systems*, Santorini, Greece, May 2008.

[30] D. H. Klatt and Klatt L. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[31] D. Koks and S. Challa. An introduction to Bayesian and Dempster-Shafer data fusion. Technical Report AR-012-775, Defence Science and Technology Organisation, November 2005.

[32] Ludmila I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[33] A. Lemieux and M. Parizeau. Experiments on eigenfaces robustness. In *Proceeding of the International Conference on Pattern Recognition*, volume 1, pages 421–424, 2002.

[34] E. Lombard. Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx*, 37:101–119, 1911.

[35] S. Marcel, Y. Rodriguez, and G. Heusch. On the recent use of local binary patterns for face authentication. *International Journal on Image and Video Processing Special Issue on Facial Image Processing*, 2007. IDIAP-RR 06-34.

[36] A. Martinez and A. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[37] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo. Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of Voice*, 10(1):59–66, 1996.

[38] H. Misra, H. Bourlard, and V. Tyagi. New entropy based combination rules in HMM/ANN multi-stream ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, April 2003.

[39] B. Moghaddam and M-H. Yang. Gender classification with support vector machines. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 306–311, Washington, DC, USA, March 2000. IEEE Computer Society.

[40] R. B. Monsen and M. A. Engebretson. Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62(4):981–993, 1977.

[41] Y. Moses, Y. Adini, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:721–732, 1997.

[42] Y. Mouchetant-Rostaing, M. Giard, S. Bentin, P. Aguera, and J. Pernier. Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience*, 12(1):303–310, 2000.

[43] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 14, pages 841–848, 2002.

[44] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, January 1996.

[45] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice–Hall, 1975.

[46] E.S. Parris and M.J. Carey. Language independent gender identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 685–688, 1996.

[47] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[48] N. Poh and S. Bengio. Non-linear variance reduction techniques in biometric authentication. In *Workshop on Multimodal User Authentication*, 2003.

[49] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[50] M. Pronobis and M. Magimai.-Doss. Integrating audio and vision for robust automatic gender recognition. Technical Report Idiap-RR-73-2008, Idiap, November 2008.

[51] M. Pronobis and M. Magimai.-Doss. Analysis of F0 and cepstral features for robust automatic gender recognition. Technical Report Idiap, Idiap, 2009.

[52] D. Rendall, M. J. Owren, E. Weerts, and R. D. Hienz. Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *Journal of the Acoustical Society of America*, 118:1711–1718, 2005.

[53] S. Ruping. A simple method for estimating conditional probabilities for SVMs. *Lernen - Wissensentdeckung - Adaptivitat*, 2004.

[54] A. Samal, V. Subramani, and D. Marx. Analysis of sexual dimorphism in human face. *Journal of Visual Communication and Image Representation*, 18(6):453–463, 2007.

[55] C. Sanderson and K. K. Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480, 2004.

[56] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 2(3):519–524, 1987.

[57] K. Sjolander and J. Beskow. WaveSurfer - an open source speech tool. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, Bejing, China, October 2000.

[58] S. Slomka and S. Sridharan. Automatic gender identification optimised for language independence. In *Proceeding of the IEEE TENCON- Speech and Image Technologies for Computing and Telecommunications*, volume 1, pages 145–148, Brisbane, Qld., Australia, December 1997.

[59] D. R. Smith and R. D. Patterson. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America*, 118:3177–3186, 2005.

[60] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier Science, 1995.

[61] D. M. Tax, M. Breukelen, R. P. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying. *Pattern Recognition*, 33(9):1475–1485, 2000.

[62] I. R. Titze. Physiology of the female larynx. *Journal of the Acoustical Society of America*, 82:S90, 1987.

[63] I. R. Titze. Pysiological and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85(4):1699–1707, 1989.

[64] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[65] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

[66] I. B. Vapnyarskii. Lagrange multipliers. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, 2001.

[67] P. Vary and R. Martin. *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

[68] L. Walawalkar, M. Yeasin, A. M. Narasimhamurthy, and R. Sharma. Support vector learning for gender classification using audio and visual cues: A comparison. In Seong-Whan Lee and Alessandro Verri, editors, *SVM*, volume 2388 of *Lecture Notes in Computer Science*, pages 144–159. Springer, 2002.

[69] K. Wu and D. G. Childers. Gender recognition from speech. Part I: Coarse analysis. *Journal of the Acoustical Society of America*, 90:1828–1840, October 1991.

[70] S. Young, G. Evermann, and M Gales. *The HTK Book (for HTK Version 3.3)*. Cambridge University, 2005.

[71] Y-M. Zeng, Z-Y. Wu, T. Falk, and W-Y. Chan. Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 3376–3379, 2006.