VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁNÍ EMOČNÍHO STAVU Z HRANÉ A SPONTANNÍ ŘEČI

Emotion recognition from acted and spontaneous speech

DOKTORSKÁ PRÁCE

DOCTORAL THESIS

AUTOR PRÁCE Ing. HICHAM ATASSI

AUTHOR

VEDOUCÍ PRÁCE prof. Ing. ZDENĚK SMÉKAL, CSc.

SUPERVISOR

BRNO 2014

ABSTRAKT

Dizertační práce se zabývá rozpoznáním emočního stavu mluvčích z řečového signálu. Práce je rozdělena do dvou hlavních častí, první část popisuju navržené metody pro rozpoznání emočního stavu z hraných databází. V rámci této části jsou představeny výsledky rozpoznání použitím dvou různých databází s různými jazyky. Hlavními přínosy této části je detailní analýza rozsáhlé škály různých příznaků získaných z řečového signálu, návrh nových klasifikačních architektur jako je například „emoční párování“ a návrh nové metody pro mapování diskrétních emočních stavů do dvou dimenzionálního prostoru. Druhá část se zabývá rozpoznáním emočních stavů z databáze spontánní řeči, která byla získána ze záznamů hovorů z reálných call center. Poznatky z analýzy a návrhu metod rozpoznání z hrané řeči byly využity pro návrh nového systému pro rozpoznání sedmi spontánních emočních stavů. Jádrem navrženého přístupu je komplexní klasifikační architektura založena na fúzi různých systémů. Práce se dále zabývá vlivem emočního stavu mluvčího na úspěšnosti rozpoznání pohlaví a návrhem systému pro automatickou detekci úspěšných hovorů v call centrech na základě analýzy parametrů dialogu mezi účastníky telefonních hovorů.

KLÍČOVÁ SLOVA

Rozpoznání emocí, řečový signál, klasifikace, spektrální příznaky, příznaky kvality řeči, spontánní řeč, analýza dialogu, call centru, komplexní klasifikační struktury, fúze

ABSTRACT

Doctoral thesis deals with emotion recognition from speech signals. The thesis is divided into two main parts; the first part describes proposed approaches for emotion recognition using two different multilingual databases of acted emotional speech. The main contributions of this part are detailed analysis of a big set of acoustic features, new classification schemes for vocal emotion recognition such as “emotion coupling” and new method for mapping discrete emotions into two-dimensional space. The second part of this thesis is devoted to emotion recognition using multilingual databases of spontaneous emotional speech, which is based on telephone records obtained from real call centers. The knowledge gained from experiments with emotion recognition from acted speech was exploited to design a new approach for classifying seven emotional states. The core of the proposed approach is a complex classification architecture based on the fusion of different systems. The thesis also examines the influence of speaker’s emotional state on gender recognition performance and proposes system for automatic identification of successful phone calls in call center by means of dialogue features.

KEYWORDS

Emotion recognition, speech signal, classification, spectral features, perceptual features, voice quality features, spontaneous speech, dialogue analysis, call center, complex classification architectures, fusion

ATASSI, H. *Emotion recognition from acted and spontaneous speech.* Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 2014. 140 p. Supervised by prof. Ing. Zdeněk Smékal, CSc.

DECLARATION

I declare that I have elaborated my doctoral thesis on the theme of "*Emotion recognition from acted and spontaneous speech*" independently, under the supervision of the doctoral thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis.

Brno . . . .   ……………………………….

(Author's signature)

Acknowledgement

I would like to express grateful thanks to my supervisor, Prof. Zdeněk Smékal, for his encouragement, patient guidance and advice he has provided throughout my time as his bachelor's, master's and doctoral student. I consider him my life mentor who taught me many things, not only technically, but also personally.

I would like to thank Prof. Kamil Vrba, the former chief of the Department of Telecommunications at Brno University of Technology for giving me the opportunity to participate in several interesting research projects.

My special thanks to Prof. Anna Esposito, International Institute of Advanced Scientific Studies, Italy and Prof. Amir Hussain, University of Stirling, UK for valuable advice and guidance given to me during my stay at their institutes.

I would also like to thank my family for their love, understanding, encouragement and care throughout past years.

Lastly, I would like to devote this thesis to all my friends, relatives and neighbors who passed away in Syria. May god brings peace to their souls and to all human beings around the world.

SIX
research centre
sensor, information and communication systems

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic
http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této doktorské práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno ....
(podpis autora)

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

2007–13
OP Výzkum a vývoj
pro inovace

Contents

List of Figures

List of Tables

List of Abbreviations

4HzME	4Hz Modulation Energy
BDES	Berlin Database of Emotional Speech
ANN	Artificial Neural Network
BS	Backward Selection
CVA	Cumulative Voice Activity
CPP	Cepstral Prominence Peak
DT	Decision Tree
DC	Direct Current
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EM	Expectation Maximization
FFBP	Feed Forward Back Propagation
FFT	Fast Fourier Transform
FS	Forward Selection
GMM	Gaussian Mixture Model
HFCC	Human-Factor Cepstral Coefficients
HMM	Hidden Markov Model
HP	High Pass filter
IDES	Italian Database of Emotional Speech
IDFT	Inverse Discrete Fourier Transform
k-NN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LFBS	Linear frequency Bank Spectrum
LFCC	Linear Frequency Cepstral Coefficients
LP	Low Pass filter
LPC	Linear Prediction Coding/Coefficients
MAE	Mean Absolute Error
MCME	Mel Cepstrum Modulation Energy
MELBS	Mel Frequency Bank Spectrum
MFCC	Mel Frequency Cepstral Coefficients
mRMR	minimum Redundancy Maximum Relevance
MSE	Mean Square Error
MSDES	Multilingual Spontaneous Database of Emotional Speech
NBC	Naïve Bayesian Classifier
PCA	Principal Component Analysis
PCBF	Perceptual Critical Band Filtering
PLP	Perceptual Linear Predictive

RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SFFS	Sequential Floating Forward Selection
SVM	Support Vector Machines
SVR	Support Vector Regression
TE	Temporal Energy
TEO	Teager Energy Operator
TMV	Temporal Mean Vector
VAD	Voice Activity Detection/Detector
VQ	Vector Quantization
VQF	Voice Quality Features
WADE	Wavelet Decomposition
ZCR	Zero Crossing Ratio

1 Introduction

In these days digital signal processing has become an indispensable part of many scientific and technical fields, such as telecommunications, robotics, biomedical engineering and biology among many others. Despite that all signals in nature are in analogue form, the computers gave us the opportunity to think about better form of processing, by converting the analogue signals to digital using analogue-to-digital converters and subsequently processed them on computers or signal processors. Digital signal processing is science that deals with mathematical methods for signal restoration, reinterpretation, enhancement and compression. It does not deal with how to understand the content of signals. For example, in speech processing, the signals can be processed to reduce the noise or to reduce the transmission speed. The methods and techniques used for understanding the signal content are a part of artificial intelligence science and its subtasks called machine learning and pattern recognition. If we go back to the example mentioned above regarding speech signals, the combination of digital signal processing and machine learning techniques can result in identifying several components, such as speaker's identity, gender, age and emotional state. Speech communication as the most natural way of inter-human understanding and interaction was not left uninvestigated from the engineers' side. This topic has been a target for research in the last few decades and a joint effort has been made from engineers, phoneticians and psychologists with the aim to better understand the nature of humans' speech.

Emotion can be defined as a conscious experience characterized by physiological expression, mental states and biological reactions. Picard in her outstanding book on affective computing [Pic00] mentioned that the vast majority of theories on emotions can be examined in terms of two main components: 1- emotions are cognitive emphasizing their mental component and 2- emotions are physical, emphasizing their bodily component. Whereas the cognitive component focuses on understanding situations that evoke emotions, the physical component deals with the physiological response that occurs when expressing emotional states.

The emotional state of a speaker can dramatically change the meaning of the speech content, for example, saying sentence "you came early today" in a sarcastic way completely change the meaning of this sentence, and gives the listener information that the speaker is not satisfied. The importance of human emotions can be proven by the fact that emotions in speech can be recognized by young children even before they can understand the speech content itself.

Automatic recognition of human emotions has been a hot topic of research over the last decade, with a vast number of published papers exploiting several modalities, such as face [GS05, PP06], physiological signals including electrocardiogram (ECG) and Electroencephalography (EEG) [ACS09, Koe+10], gestures [CKC08], postures [Cou04], pupillary dilation [PS03] and speech. The last mentioned modality can be considered as one of the most targeted in terms of human emotion recognition (the first paper on this topic was published in 1984 [Van84]). The interest of engineers in developing systems capable of recognizing human's emotions from speech can be explained by the unlimited applications of such systems in many fields, such as robotics, telecommunications services, e-learning among many others.

A very important utilization for vocal emotion recognition can be found in call centers, which are centralized organizations aiming to receive or transmit a big amount of requests by telephone. The aim of call centers is to handle phone orders, customer support, emergency services and telemarketing. The call centers have become a very important part of world economy especially in the past 10 years. In 2001, around 3% of labor force in The US and Canada has been working at a call center [AL05]. The analysis of telephone conversations in call centers is gaining importance since the companies working in this field figured out that the evaluation of performance of operators in such companies is crucial as well as the clients' feedback. However, it is very hard to manually evaluate the quality of services provided, or to assess the agents' performance. For example, if a company has 20 operators working daily for 7 hours and 5 days per week, then the phone calls recorded throughout one month make about 2800 hours. Obviously, it is impossible to manually check all these phone calls in order to make a reliable image about agents' performance or to assess the quality of services. In the light of mentioned above, it appears evident that a kind of automatic analysis telephone records is indispensable for phone-marketing and all subjects that involve costumer support services in their structure.

The thesis is organized as follows

- **Chapter 2** gives a brief overview of literature related to vocal emotion recognition.
- **Chapter 3** describes in details acoustic features that might be used for emotion recognition, with a special focus on prosodic and perceptual spectral features.
- **Chapter 4** gives an overview of feature section methods for the identification of features with best discriminative power in terms of distinguishing between emotional states.
- **Chapter 5** deals with methods used for emotion classification.

- **Chapter 6** describes proposed approaches for emotion recognition using two different databases of acted emotional speech. The main contributions of this part are detailed analysis of a big set of acoustic features and new classification schemes for vocal emotion recognition such as “emotion coupling”.
- **Chapter 7** presents new multilingual spontaneous database of emotional speech based on telephone records from real call centers and the corresponding approaches for emotion recognition from this database.
- **Chapter 8** studies the influence of speaker’s emotional state on gender recognition performance.
- **Chapter 9** presents novel system for automatic identification of successful phone calls in call center by means of dialogue features.
- **Chapter 10** contains conclusions and final remarks.

The thesis aims can be summarized as follows

- Propose new approaches for vocal emotion recognition from acted speech databases by means of complex classification architectures.
- Exhaustive analysis of a big set of speech features that might be used for vocal emotion recognition.
- Design autonomous system for emotion recognition from spontaneous speech.
- Propose a method for mapping outputs of discrete systems for vocal emotion recognition into continuous two-dimensional space
- Study the influence speaker’s emotional state on the performance of gender recognition.
- Propose a method for automatic identification of successful phone calls in call center by means of dialogue features.

2 State of the art

Emotion recognition from speech can be approached, in general, as any pattern recognition task. Figure 2.1 illustrates basic scheme of emotion recognition system followed by the literature overview.

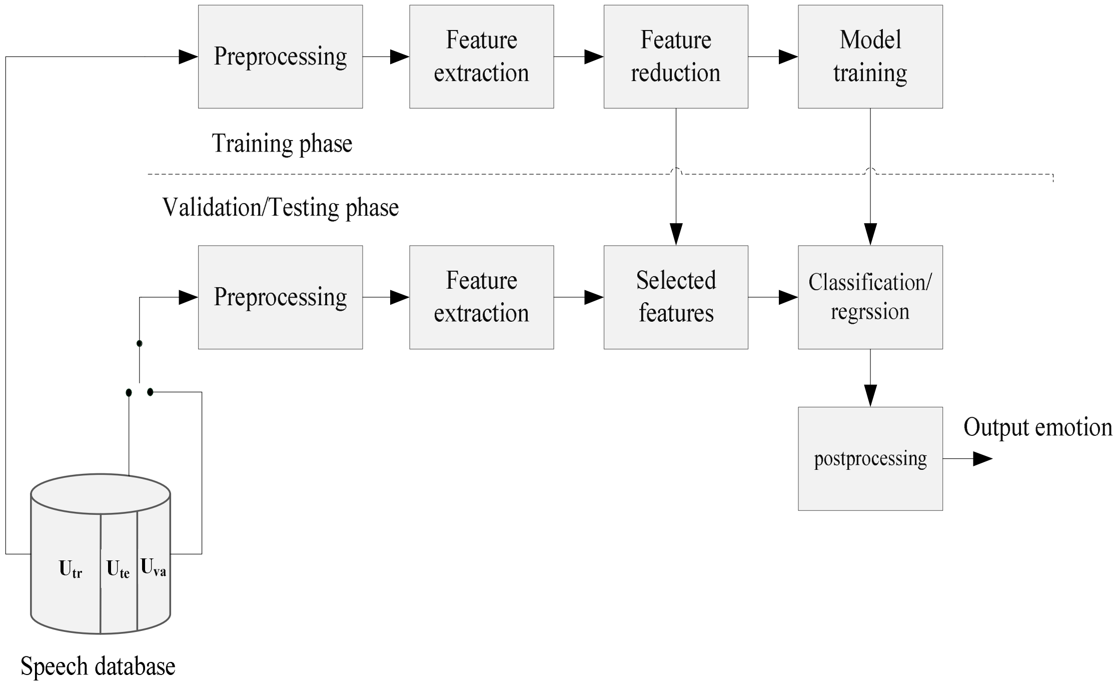


Figure 2.1: Basic scheme (life cycle) of emotion recognition from speech (U_{tr},U_{te} and U_{va} are groups of utterances for training, testing and validation respectively).

2.1 Speech databases

Having a well constructed speech database is a crucial issue before starting any experiments with speech processing, and the emotion recognition is not an exception. The utterances of this database are used to train the decision making model (classification or regression algorithm) that is subsequently used to classify testing (hidden) patterns. There are three types of emotional speech material usually used for vocal emotion recognition.

2.1.1 Acted speech

Speech records are commonly obtained from normal people or actors, where the speakers are asked to produce short sentences with different emotions. Such databases are in most cases recorded in a studio, which guarantees a high quality material. However, several studies showed that acted speech recorded in studio conditions is not authentic in terms of expressed emotions. Nevertheless, the acquisition of acted emotional speech is easy in comparison with the other types of emotional speech databases. The most used database of acted speech is Berlin Database of Emotional speech [Bur+05]. Other examples of this type of databases include Sweden Emotion Speech Database [AA00] and Belfast Database [Dou+00].

An alternative to acted emotional speech in studio conditions, where protagonists are usually asked to stand in front of a microphone and express predefined emotions is using speech extracts from movies or TV series. Distinct from the existing emotional databases, this kind possesses a certain degree of spontaneity since the actors/actresses producing the sentences are acting according to the movie script and their performance are judged as appropriate to the required emotional context by the movie director (supposed to be an expert). As examples of these kind of databases there are COST2102 Database of Emotional speech [ERM09] and Microsoft Research China Recorded Database [Yu+01].

2.1.2 Elicited speech

This emotional material is acquired using the following procedure: each speaker is firstly asked to see a film or to read a short paragraph. After that a kind of spontaneous discussion is held on between the subject and the person responsible for the test. The authenticity of elicited speech is higher in comparison with acted speech. Moreover, it is possible to obtain good quality records as the sessions can be carried out in a quiet place. Several attempts have been made so far to construct elicited speech databases, such as in [NNT00, SHC10].

2.1.3 Spontaneous speech

Emotions spontaneously expressed have evidently the highest authenticity. However, it is very hard to collect such material with a fair audio quality. Another problem appears from the ethical point of view, where it is not acceptable to record people without their knowledge or to induce emotions by telling people fake information. Lost Luggage Study Corpus [Sche00] presents one of the early attempts to create speech databases with spontaneous emotions.

In General, it can be stated that the performance of approaches for automatic vocal emotion recognition are negatively correlated with the authenticity of speech content. For example Pfister and Robinson proposed in their paper [PR11] real-time approach for emotion recognition for public speaking skill analysis. Spontaneous speech material from Mind Reading corpus was used for experiments. Support vector machine classifier with MFCC, MELBS and ZCR based suprasegmental features were employed to recognize 5 emotional states. The average classification accuracy was 70%. On the other hand, it is very common to reach very high classification accuracies, such as in [KL06], where 90% classification accuracy was achieved on acted Mandarin speech by using MFCC, energy, fundamental frequency with SVM classifier.

It should be noticed that the performance of methods for emotion recognition are not comparable when different emotional databases are used, since the acoustic features of emotional speech vary primarily according to the language as well as to the recording environments, the effectiveness of the subjects involved in the recordings as well as if the emotional state is acted or spontaneous [VE05]. Figure 2.2 illustrates the relation between the authenticity of content and the complexity of acquirement for acted, elicited and spontaneous speech

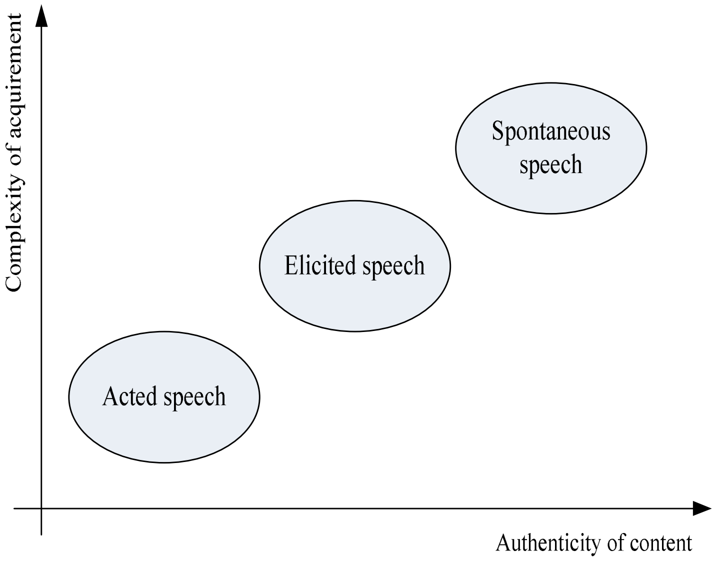


Figure 2.2: The relation between the authenticity of content and the complexity of acquirement for acted, elicited and spontaneous speech.

2.2 Overview of methods for feature extraction, selection and classification

Different types of features were employed in literature for vocal emotion recognition. The most known among them are prosodic and voice quality features including fundamental frequency [NFT06, BLN09, Ste+08], temporal energy [SRL04, CS07], Teager Energy operator [LZZ10, ZHK01, Formant frequencies [CK11, Pet00, Vla11], Zero Crossing Rate [RLC09] and voice-source parameters [Sun+11]. Several studies employed spectral-shape features [EGP10, MVG10, VK06], LPC [SFP09] and LPCC [Nwe+03]. The perceptual spectral features were also successfully exploited for emotion recognition from speech. These features include MFCC [BGC06, SFP09], MELBS [BLS07, AS08] and PLP [SSP09, AAE00]. Other rarely used features for vocal emotion recognition are based on wavelet transform [GLC03].

The feature reduction is very important part of pattern recognition systems which aims to find features with best discriminative power. Feature reduction methods can be divided into two main groups: methods for feature transformation such as *k*-means used in [SRL04, SFP09], PCA [EGP10, MN11] and VQ [SRL04]. The second group contains methods for feature selection, such as Relief algorithm exploited in [Pet00], wrapping techniques [VK06] and simple methods based on t-test.

Automatic recognition of emotional vocal expressions could be carried out by using different kinds of classifiers, such as Artificial Neural Networks [Sch+08, Stu+11, NNT00, BGC06], Fuzzy logic systems [LN03, Cha+09], Hidden Markov Models [SRL03, ZPR09], *k*-Nearest Neighbour [AHS11,,PCY04, Pao+07, SSP09], Support Vector Machines [SRL04, Yac+03, EGP10] Gaussian Mixture Models [SRL04, VK05] and Decision Trees [CS07, SY12].

Some research employed more complex approach to the classification of vocal emotions. For example, paper [WL11] presents a fusion-based approach by using several classifiers in combination with acoustic prosodic features and semantic labels. The results showed that combining acoustic information and semantic labels can achieve 83.55% accuracy for speaker-dependent approach and 85.79% for speaker-independent approach.

The work presented in [Lef+10] explores possibilities for enhancing portability, generality and robustness of vocal emotion recognition by combining several databases and by fusing different classifiers. Another work considering the fusion of multiple classifiers was reported in [ZPR09], where Hidden Markov Model and Neural Network were combined aiming to enhance the classification accuracy of human emotions. The work presented in [SFP09] examines the possibility of combining several Radial Basis Neural Networks on different sets of features, including MFCC, LPC, energy and F0.

The paper presents in [KL12] provides an empirical study of the most widely used classifiers in the domain of emotion recognition from speech, across multiple non-acted emotional speech corpora. The results indicate that Support Vector Machines have the best performance and that they along with Multi-Layer Perceptron networks and *k*-nearest neighbor classifiers perform significantly better than decision trees, Naive Bayes classifiers and Radial Basis Function networks.

The paper presented in [Kim+10] proposed a method of recognizing emotion in speech with a dimensional approach by combining semantic and acoustic features. The training corpus contains movie clips of audio and text in a subtitle format rated in arousal and valence dimensions by human subjects. The authors showed that combining the semantic and acoustic features improve the recognition results and also showed that semantic features are better at estimating the valence dimension while the acoustic features are better at eliciting the arousal dimension.

There are two key factors that should be taken into account in terms of emotion recognition from speech

1. **Speaker's gender**
   Based on the work introduced in [YP11, Kim09], it can be stated that gender-dependent systems achieve better classification accuracy of emotional states comparing to gender-independent systems. This approach doesn't pose any serious limitation of functionality since the recognition of speaker's gender can be carried out with high accuracy.

2. **Speaker dependency**
   In terms of speaker dependency there are two types of systems for vocal emotion recognition:
   - *Speaker-dependent systems*: When utterances of one speaker are used for training it is possible to use utterance of the same speaker for testing.
   - *Speaker-independent systems*: The opposite of the previous case; for speaker independent systems it is not allowed to use utterances of the same speaker for both training and testing.

Based on survey of research on vocal emotion recognition, it can be concluded that speaker-dependent systems, being less complex, provides better performance and gives much better results than speaker-dependent systems as it is reported by Navas et al. [NHL06] (98% classification accuracy for seven emotions by using a GMM classifier). When the task aims to be speaker-independent the better results achieved were around 75% as reported in the work of Lugger & Yang [LY07].

2.3 Overview of research on call centers

The research in the area of call centers usually explores the simulation and modeling of these structures [AL05, AW13, Con13]. Only a small fraction of the research has been devoted to the automatic analysis of recorded telephone calls through speech processing and machine learning. The next is devoted to give an overview of some attempts to bring the last two mentioned disciplines into the domain of call centers.

In [AK12], the authors achieved encouraging results in identifying potential problem calls in call centers by using only speaking rate as a feature. However, the proposed approach is semi-autonomous and requires human supervision. On the other hand, the work described in [TPH03] is fully autonomous and presents one of the early studies on applying machine learning to the domain of call centers. The authors proposed a method for call-type classification in the context of telephone-based speech corpus. The method presented in the mentioned work is able to automatically classify calls into 6 classes and considers the outputs of both acoustic and language models for decision making. The Support Vector Machine (SVM) classifier provided the best classification accuracy with error rate of 24% for manually transcribed conversations. Another study bringing machine learning to the domain of call centers was published in [JM10], the authors of this work proposed an improved neural network algorithm for the prediction of call center service grade.

Emotion recognition from telephone records has also been an area of focus in several studies; Authors in [YP07] introduced a speech emotion recognition system for call centers. The system can deal with two emotional states – neutral and anger from speech captured by a cellphone in real time. The effect of environmental noise was alleviated by using a moving average filter on the feature space. The speech data used for experiments includes short utterances by fifteen semi-professional actors and it consists of total of 5400 utterances for both mentioned emotional states. Another work that was introduced for call center application is [Yac+03]. The aim of this work was to distinguish anger from neutral speech which was achieved with 91% accuracy.

The aim of research presented in [VS11] was to find a way to automatically detect the customer's emotional state and to signal when the discontentment level reaches a critical value. Since the classification of four emotions (neutral, nervous, querulous, and others) was not successful, the emotions were fused into one class and classified against the neutral state. The average classification accuracy in this case was 73%. This result was achieved by using MFCC and F0 as features and SVM classifier. Authors in [Lau+11] utilized a large corpus of spontaneous emotional speech recorded from real-life voice controlled telephone services. Listeners rated a selection of 200 utterances from this corpus with regard to level of perceived irritation, resignation, neutrality, and emotion intensity. The features extracted were F0, intensity, formants, voice-source parameters and temporal characteristics of speech. As a classifier, the authors employed linear discriminant Analysis to recognize irritation, resignation, and neutral. The classification accuracy of these three states was about 62%.

In [CD11], the detection of spontaneous emotions is explored using different speech corpora with only two emotional states: anger and neutral. The speech corpora were collected in call centers in different contexts (service complaints, stock exchange service and medical emergency). The usage of several speech corpora with different contexts aimed to find the influence of the context on the classification accuracy. The proposed system was based on the extraction of F0, ZCR, MFCC and energy as features and SVM classifier. The best classification accuracy of anger/neutral among contexts was 85.3%.

3 Feature extraction

This chapter is devoted to the description of features that might be used for emotion recognition from speech. In general, such features can be categorized into three main groups: Prosodic features, spectral features and voice quality features. Beside these mentioned feature groups, some other features were employed in experiments presented in this thesis and will be also briefly introduced in this chapter.

Before the feature extraction, the speech signal should be processed in several steps before it can be passed to further processing. The main objective of preprocessing stage is to enhance and adapt the speech signal to the feature extraction process. The preprocessing commonly includes the following steps

1. **Resampling**: in many cases, the input speech signal might have a high sampling frequency (over 16 kHz). However, such frequency is higher than enough to represent the entire frequency spectrum, thus it is recommended to remove the redundancy by down sampling the speech signal.
2. **Removing of the direct current (DC) component**: The DC component doesn't carry any meaningful information and moreover it can result in bad extraction of some features such as the temporal energy. The DC component is removed from speech signal by the following simple equation

   $$s[n] = s[n] - \tfrac{1}{L}\textstyle\sum_{i=0}^{L-1} s[n], \tag{3.1}$$

   where $s[n]$ is the speech signal and L is number of its samples.

   In some applications of speech processing, it is also appropriate to remove silence parts of speech by using voice activity detection algorithm.

3. **Normalization**: It is important to have the signal normalized in terms of its dynamic to the range of -1 to +1 since most of the features extracted from normalized and non-normalized signal can significantly differ. The normalization is done by the following formula

$$s[n] = \frac{s[n]}{\max(|s[n]|)}. \tag{3.2}$$

4. **Pre-emphasis**: This processing is applied to boost the magnitude of higher frequencies aiming to flatter the magnitude spectrum and to balance the magnitude of high and low frequencies. The pre-emphasis is usually carried out by simple first-order FIR filter defined by the following transfer function

   $$H(z) = 1 - \alpha z^{-1}, \quad \alpha \in [0.9,1]. \tag{3.3}$$

5. **Segmentation and windowing**: Since speech is non-stationary signal with quick changes of features over time, it is in most cases processed after segmenting into short frames. These frames are subsequently multiplied by windowing functions. Despite the existence of various windowing functions in speech processing domain with different effects on the spectrum shape of the windowed frame, we didn't observe any significant impact of the type of window function on the classification performance of emotions; hence the Hamming window is used in all experiments presented in this thesis. The segmentation process and the preprocessing methods are illustrated in Figures 3.1 and 3.2 respectively.

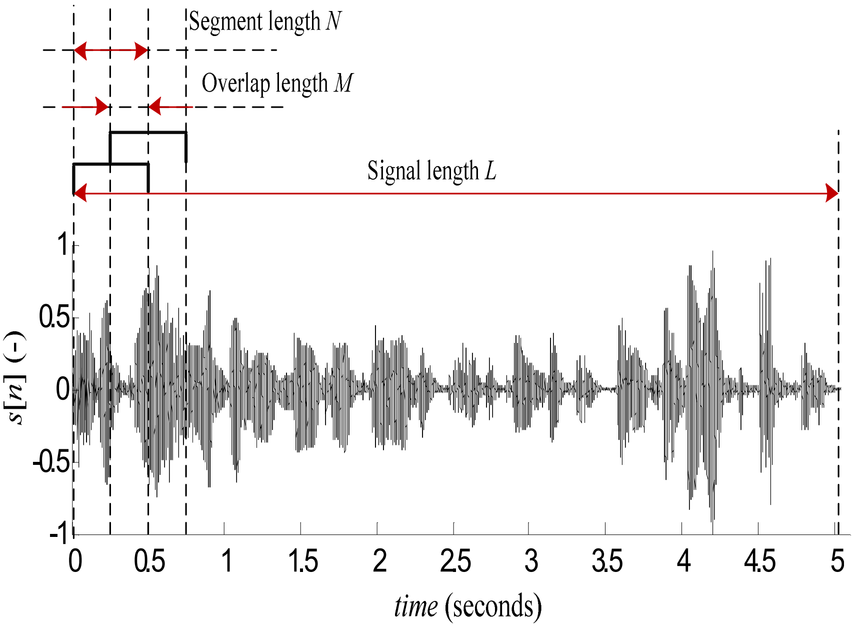


Figure 3.1: Illustration of segmentation process of speech signal.

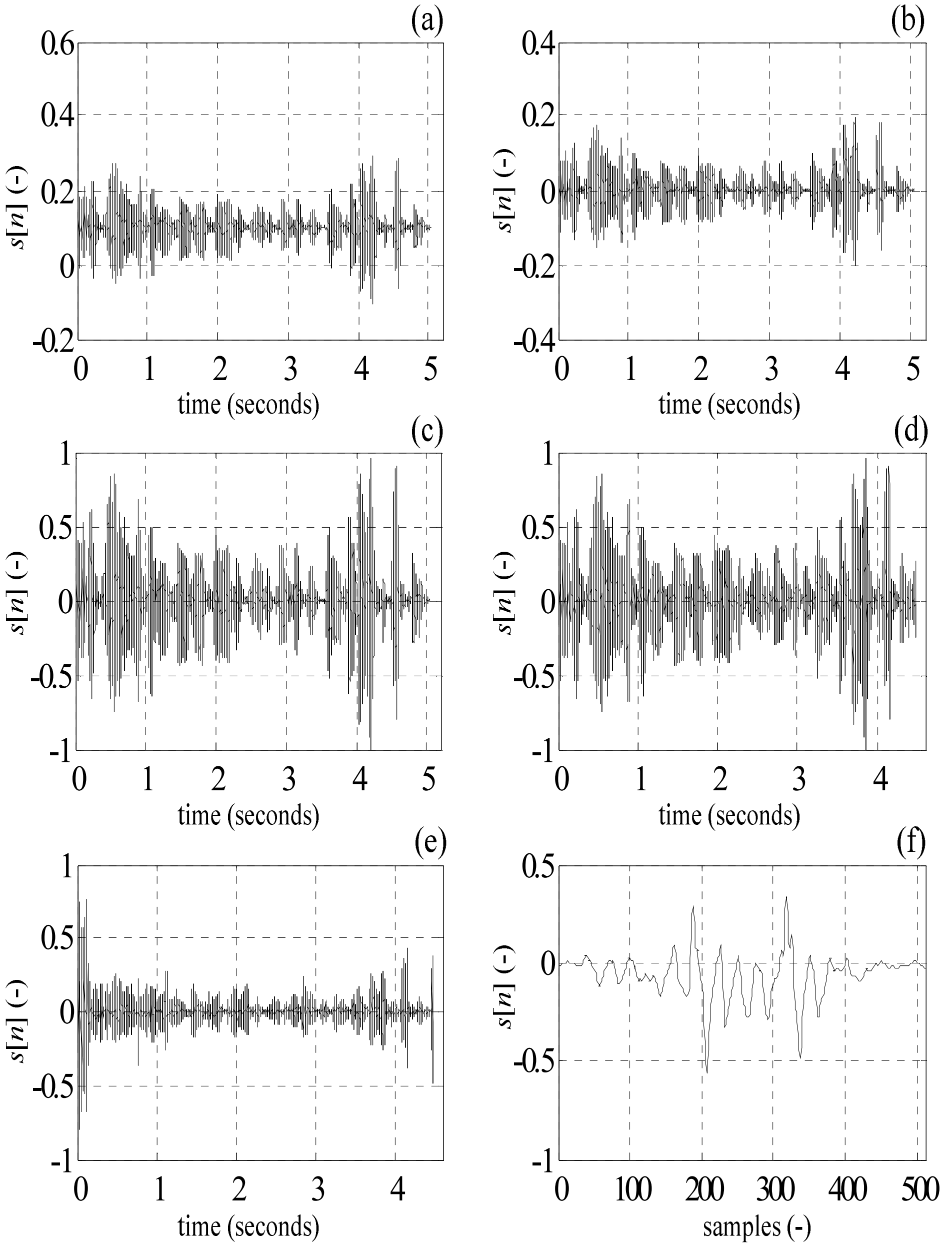


Figure 3.2: Illustration of preprocessing methods: (a) Input speech signal (b) signal after removing DC component (c) signal after normalization (d) Signal after silence removing (e) signal after pre-emphasis (f) 512 samples long segment extracted from the original signal by applying Hamming window.

In the next sections, the features used in experiments with emotion recognition from speech are briefly presented.

3.1 Prosodic features

These features were primarily addressed by phoneticians, especially the fundamental frequency, which is the main carrier of information about speech intonation. According to own experiments, it was observed that the usage of the entire pitch waveform doesn't bring a significant improvement from the classification accuracy point of view, especially when this waveform is used with static classifiers such as GMM or FFBP-ANN. On the other hand, the pitch or temporal energy waveforms are frequently combined with HMM as this classifier includes the temporal variation of features in the classification process.

The next will be devoted to the most known prosodic features used for emotion recognition

3.1.1 Fundamental frequency

This feature is also known as pitch or F_0. The fundamental frequency was considered as the most relevant feature for vocal emotion recognition for a long time. There are several methods proposed for speech fundamental frequency estimation [Ata08], the most known among them are based on autocorrelation function.

3.1.2 Temporal energy

This feature is one of the oldest features used in speech processing; the temporal energy has several mathematical definitions, the most known is given as

$$E[m] = \sum_{i=0}^{N-1} (s[n])^2, \tag{3.4}$$

where $E[m]$ is the temporal energy of the m-th speech segment $s[n]$ and N is the segment length. Another feature that represents the signal energy is the Teager energy operator (TEO) [Kai90], which characterizes the signal energy in a nonlinear manner.

3.1.3 Zero crossing rate

This feature is defined as the number of sign changes along the speech signal. Although its mathematical interpretation is simple, it might be a simple alternative of fundamental frequency. The ZCR is given by the following formula

$$Z[m] = \frac{1}{M}\sum_{n=0}^{M-1} |\mathrm{sign}(s[n]) - \mathrm{sign}(s[n-1])|. \tag{3.5}$$

3.2 Speech Quality features

These features are usually employed for the detection of pathological and neurological disorders [MRS11]. However, the importance of such features for emotion recognition from speech is obvious from many research papers. Some theories suggest that speech quality features (SQF) play a key role in discriminating between positive and negative emotions, such as anger and happiness.

3.2.1 Harmonicity

It is quite difficult to make a straight definition of this feature due to the variety of implementations observed in relevant literature. In [Boe02] the author defines harmonicity as the representative of the degree of acoustic periodicity in speech signal. In own experiments, the harmonicity is defined as the ratio given in dB between the first maximum of autocorrelation function and signal energy. The harmonicity of the m-th speech segment with maximum autocorrelation $R_{\max}$ and energy represented by the first coefficient of the autocorrelation function R_0 is given as follows

$$H[m] = 20\log(\frac{R_{\max}}{R_0}). \tag{3.6}$$

3.2.2 Formants

Formants are frequency peaks with high energy in the spectrum especially prominent in vowels and represents resonances in the vocal tract. There are several approaches proposed in literature for formant extraction. In this thesis, the method based on linear prediction spectra was used, which was firstly proposed in [Yeg78].

3.2.3 Cepstral prominence peak

Cepstral prominence peak firstly proposed in [HSG02] is a measure of the amplitude of the cepstral peak that corresponds to the fundamental frequency, normalized for all signal amplitude. This feature contains information about the degree of harmonic organization in speech signal.

3.3 Spectral features

The spectral features considered in this thesis are divided into two main groups: Basic spectral features and Perceptual spectral features.

3.3.1 Basic spectral features

These features are widely employed in speech and music processing domain as they capture the changes of spectrum in time. Four different spectral characteristics are employed in own experiments:

1. Spectral centroid

$$S_{\mu} = \frac{\sum_{n=0}^{N-1} f[k]S[k]}{\sum_{k=0}^{N-1} S[k]}, \tag{3.7}$$

2. Spectral spread

$$S_{\mathrm{s}} = \frac{\sum_{n=0}^{N-1}\left(f[k] - S_{\mu}\right)^{2} S[k]}{\sum_{k=0}^{N-1} S[k]}, \tag{3.8}$$

3. Spectral skewness

$$S_{\sigma} = \frac{\sum_{n=0}^{N-1}\left(f[k] - S_{\mu}\right)^{3} S[k]}{\sum_{k=0}^{N-1} S[k]}, \tag{3.9}$$

4. Spectral kurtosis

$$S_{\gamma} = \frac{\sum_{n=0}^{N-1}\left(f[k] - S_{\mu}\right)^{4} S[k]}{\sum_{k=0}^{N-1} M[k]}, \tag{3.10}$$

5. Spectral slope

$$S_{\mathrm{sl}} = \frac{\sum_{k=0}^{N-1} f[k]S[k] - \sum_{k=0}^{N-1} f[k] \sum_{k=0}^{N-1} S[k]}{\left[\sum_{k=0}^{N-1} f^{2}[k] - \left(\sum_{k=0}^{N-1} f^{2}[k]\right)^{2}\right] \sum_{k=0}^{N-1} S[k]}, \tag{3.11}$$

Where $S[k]$ is the DFT/FFT module in dB and $f[k]$ is the frequency in Hz.

3.3.2 Perceptual spectral features

All features belonging to this group are extracted from the speech DFT/FFT module. They differ in two things: The type of filter bank used to reduce the dimension of DFT/FFT module and the further processing. The perceptual spectral features considered in this work are

1. **Mel-Frequency Cepstral Coefficients (MFCC)**: These well-known features are adopted, as an encoding method, in many fields of speech and audio signal processing among those musical genre recognition, speaker and speech recognition. And in many dedicated works on the recognition of emotional vocal expression.
2. **Linear-Frequency Cepstral (LFCC)**: these features are calculated analogously to MFCC but a filter bank with linear scale is used instead of mel bank.
3. **Human Factor Cepstral Coefficients (HFCC)**: Firstly proposed in [SH03] as a better noise-robust alternative of MFCC for speech recognition.
4. **Mel Bank Spectrum (MELBS)**: Are taken by multiplying the DFT/FFT module of a speech segment by mel filter bank. These features are rarely used for emotion recognition, one of the exceptions is the work presented in [28]. MELBs are employed in the speech processing domain as features for Voice Activity Detector algorithms (VAD) [ESS07] as well as for speech segmentation [EA05]. However it should be mentioned here that many proposed papers have employed spectral energies in different frequency bands as features to recognize emotions from speech, and analogously, MELBs could be considered as an enhanced alternative for those band energies since it involves the mapping of the DFT/FFT module onto the perceptual MEL frequency scale.
5. **Linear Frequency Bank Spectrum (LFBS):** Derived analogously to MELBS but it uses linear frequency bank instead of mel bank.
6. **Perceptual Linear Predictive (PLP):** this feature was primarily introduced in [Her90] and showed better results in comparison with the typical MFCC encoding in terms of speech recogniton. In this thesis, four intermediate results of PLP extraction process were considered as independent features. The first one, PLP1 is simply the Bark spectrum of a speech segment. PLP2 is the post-processed Bark spectrum, where PLP1 is multiplied by the equal loudness curve and the module is compressed. The third PLP type is the LPC smoothed spectrum of PLP2 and the last type PLP4, is the smoothed LPC cepstrum of PLP2.

The block scheme of perceptual feature extraction process is illustrated in Figure 3.3 and Figure 3.4 shows the width of the filters used within the scheme of perceptual spectral feature computation.

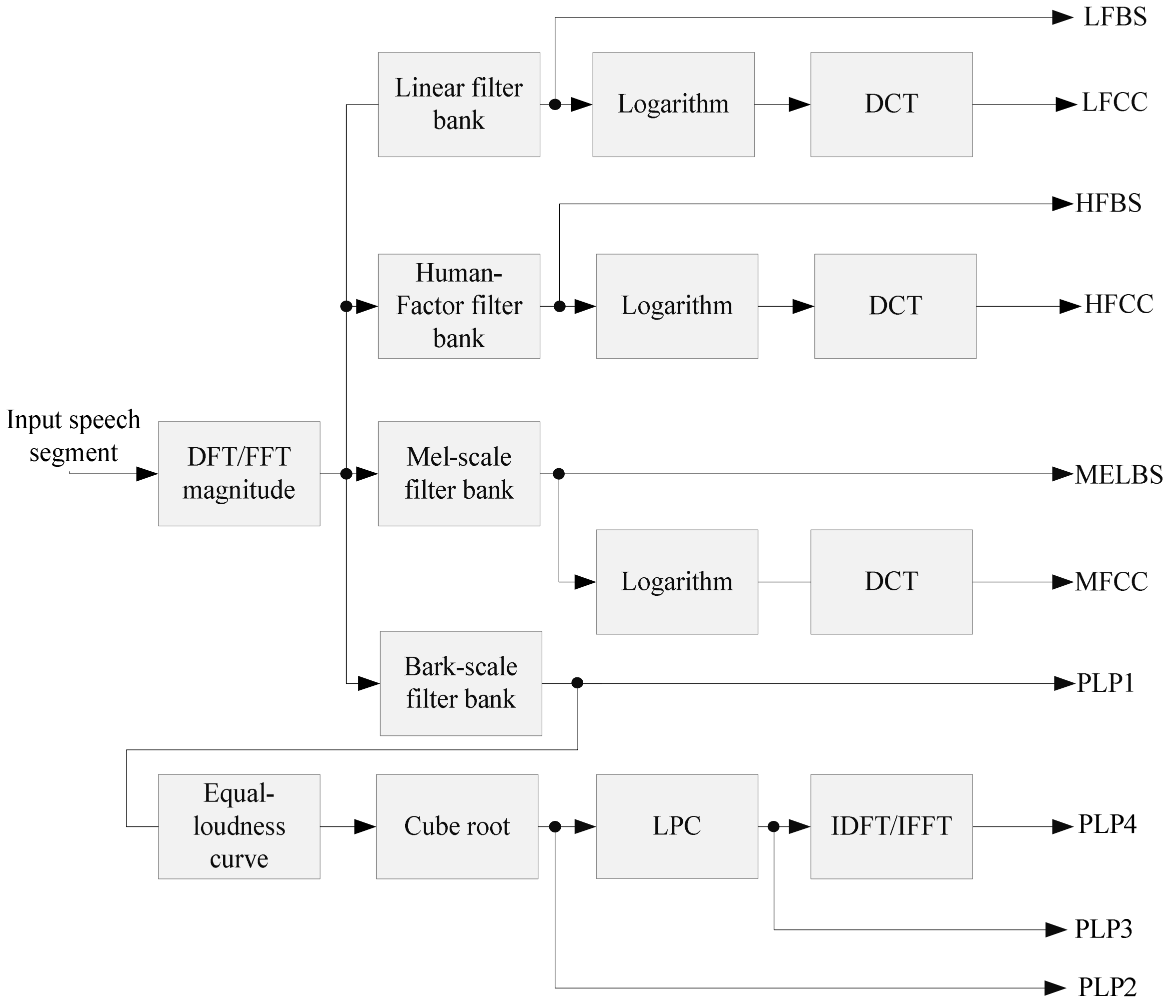


Figure 3.3: Block scheme of perceptual spectral feature extraction process.

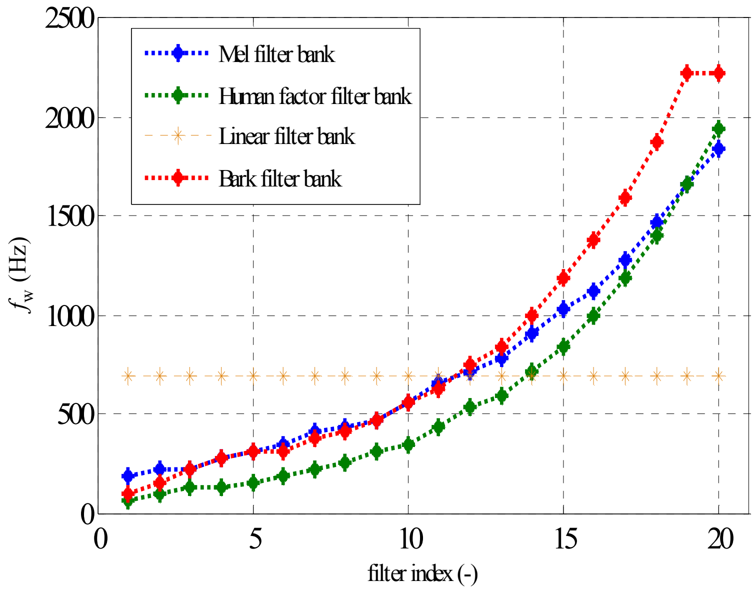


Figure 3.4: Filter bandwidths for different perceptual filter banks.

3.4 Wavelet transform based features

The wavelet transform was successfully employed in different domains of speech processing but rarely for emotion recognition. In this thesis, two types of wavelet transform based features are employed. The first one is called Subband Based Cepstral coefficients (SBC), these coefficients are extracted by wavelet packet decomposition of the frequency range 0 to 4 kHz such that the obtained 24 frequency subbands follow the Mel scale [SH00, MGS07].

I proposed another type of wavelet transform based features. The idea is to apply a symmetric three-level quadrature filter bank with Haar filters on the speech signal and subsequently extract spectral features, temporal energy and zero crossing ratio from each band. The aim of such approach is to achieve more detailed representation of speech signal. The block scheme of this feature extraction scheme is illustrated in Figure 3.5.

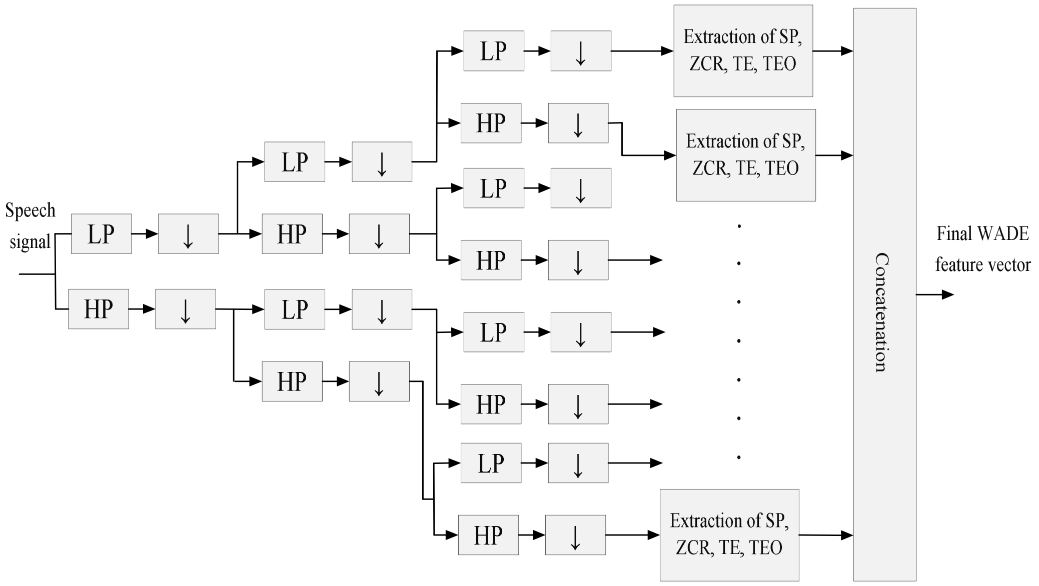


Figure 3.5: Block scheme of feature extraction based on wavelet decomposition.

3.5 modulation energy based Features

Despite this kind of features has never been exploited for vocal emotion recognition; they might be useful for this purpose as the 4Hz modulation energy represents the syllabic rate of speech [HH73]. In own experiments, I used two representatives of speech modulation energy; the first one is the simple 4 Hz modulation energy extracted by applying a LP filter, and the second one called Mel-Cepstrum Modulation Energy (MCME) is extracted from the time trajectory of Mel-frequency cepstral coefficients (MFCC) [BDY07].

Beside the features mentioned in this chapter, I tested in one experiment, as it will be shown in chapter 6. well-known Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients and Adaptive Component Weighting Coefficients [MXR96].

3.6 Suprasegmental features

Suprasegmental features, sometimes called high-level characteristics can be extracted from any feature mentioned in this chapter. Based on own experiments, these features show an excellent performance in terms of vocal emotion recognition. For an arbitrary feature vector $F[n]$ with length of N, the suprasegmental features in Table 3.1 are considered in further experiments. As an example, Figure 3.6 illustrates temporal energy extracted from speech signal and it corresponding suprasegmental features.

Table 3.1: List of suprasegmental features.

Mean	$h_{\text{mean}} = \frac{\sum_{n=0}^{N-1} F[n]}{N}$
Median	$y_{\text{median}} = \text{median}(F[n])$
Standard deviation	$h_{\text{std}} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(F[n] - \frac{\sum_{n=0}^{N-1} F[n]}{N}\right)^2}$
Relative standard deviation	$h_{\text{relstd}} = N\frac{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(F[n] - \frac{\sum_{n=0}^{N-1} F[n]}{N}\right)^2}}{\sum_{n=0}^{N-1} F[n]}$
Maximum	$h_{\max} = \max(F[n])$
Maximum position	$h_{\text{maxpos}} = \text{argmax}(F[n])$
Minimum	$h_{\min} = \min(F[n])$
Minimum position	$h_{\text{minpos}} = \text{argmin}(F[n])$
Range	$h_{\text{range}} = \max(F[n]) - \min(F[n])$
Relative range	$h_{\text{relran}} = N\frac{\max(F[n]) - \min(F[n])}{\sum_{n=0}^{N-1} F[n]}$
Relative maximum	$h_{\text{maxrel}} = N\frac{\max(F[n])}{\sum_{n=0}^{N-1} F[n]}$
Relative minimum	$h_{\text{minrel}} = N\frac{\min(F[n])}{\sum_{n=0}^{N-1} F[n]}$
Relative position of maximum	$h_{\text{relmaxpos}} = \frac{\text{argmax}(F[n])}{N}$

Relative position of minimum	$h_{\text{relminpos}} = \frac{\text{argmin}(F[n])}{N}$
Slope	$h_{\text{slope}} = \sum_{n=0}^{N-1} \frac{3\|F[n] - F[n-1]\|}{N}$
Skewness	$h_{\text{skew}} = \frac{\sum_{n=0}^{N-1}(F[n] - h_{\text{mean}})^3}{N h_{\text{std}}^3}$
Kurtosis	$h_{\text{kurt}} = \frac{\sum_{n=0}^{N-1}(F[n] - h_{\text{mean}})^4}{N h_{\text{std}}^4}$
k-th percentile	$h_{\text{perc}(k)} = \frac{kN}{100} + \frac{1}{2}$
Percentile range	$h_{\text{percran}} = h_{\text{perc}(99)} - h_{\text{perc}(1)}$
Regression coefficient	$h_{\text{regcoef}} = \frac{\sum_{n=0}^{N-1} nF[n] - \frac{1}{N}\sum_{n=0}^{N-1} n \sum_{n=0}^{N-1} F[n]}{\sum_{n=0}^{N-1} n^2 - \frac{1}{N}(\sum_{n=0}^{N-1} n)^2}$
Pearson skewness	$h_{\text{psc}} = \frac{h_{\text{mean}} - h_{\text{median}}}{h_{\text{std}}}$
5th moment	$h_{\text{5thmoment}} = \frac{\sum_{n=0}^{N-1}(F[n] - h_{\text{mean}})^5}{N h_{\text{std}}^5}$
6th moment	$h_{\text{6thmoment}} = \frac{\sum_{n=0}^{N-1}(F[n] - h_{\text{mean}})^6}{N h_{\text{std}}^6}$

3.7 Defining different levels of feature extraction

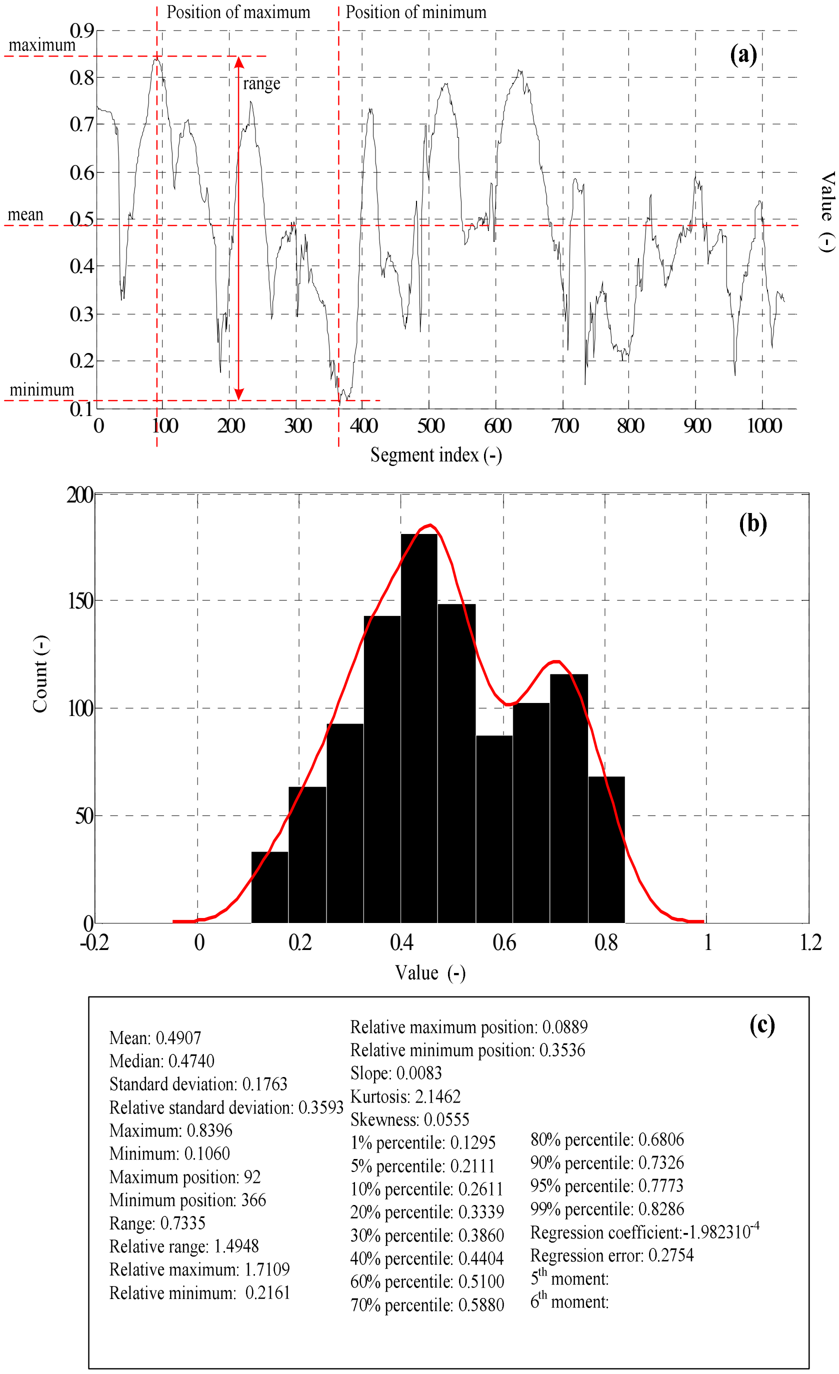
When it comes to feature extraction from speech signal, there are several levels of extraction that can be considered. Depending on the application, the selection of appropriate extraction scheme plays important role in the system performance. For instance, features for phoneme recognition are almost exclusively extracted from short segments and subsequently classified. On the other hand, for applications such as emotion recognition or social signal processing, the suprasegmental features are considered as a valuable carrier of information. In this thesis, three different levels of speech signal representation are used as it is illustrated in Figure 3.7.

The following interpretations were tested in our experiments

1. **Overall representation**: one feature vector is extracted from the supersegment. For example, if only 14 MFCC coefficients are considered then these coefficients are extracted from the entire supersegment without any segmentation. According to own early experiments, this interpretation is not useful for emotion recognition as it does not contain any temporal information.
2. **Segmental representation**: the supersegment is segmented into shorter segments from which the feature vectors are extracted. Each of these feature

vectors can be classified separately or transformed into new feature vector by using an appropriate feature transformation technique, as it will be shown in Chapter 4.

3. **Suprasegmental representation**: The suprasegmental features are extracted from the segmental feature waveform and subsequently concatenated.



Mean: 0.4907
Median: 0.4740
Standard deviation: 0.1763
Relative standard deviation: 0.3593
Maximum: 0.8396
Minimum: 0.1060
Maximum position: 92
Minimum position: 366
Range: 0.7335
Relative range: 1.4948
Relative maximum: 1.7109
Relative minimum: 0.2161

Relative maximum position: 0.0889
Relative minimum position: 0.3536
Slope: 0.0083
Kurtosis: 2.1462
Skewness: 0.0555
1% percentile: 0.1295
5% percentile: 0.2111
10% percentile: 0.2611
20% percentile: 0.3339
30% percentile: 0.3860
40% percentile: 0.4404
60% percentile: 0.5100
70% percentile: 0.5880

80% percentile: 0.6806
90% percentile: 0.7326
95% percentile: 0.7773
99% percentile: 0.8286
Regression coefficient:-1.982310^{-4}
Regression error: 0.2754
5th moment:
6th moment:

Figure 3.6: Example of suprasegmental features extracted from temporal energy waveform.

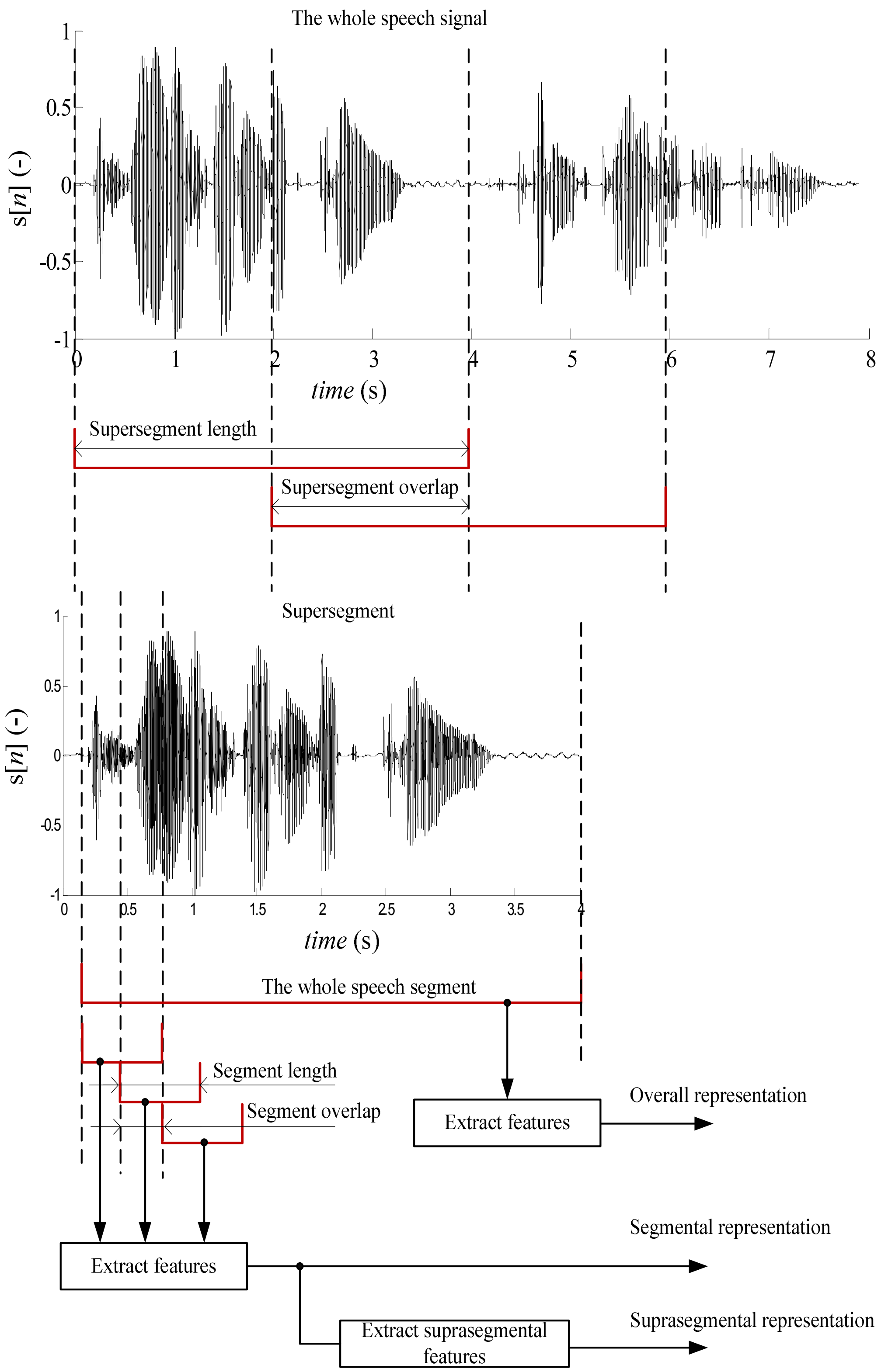


Figure 3.7: Illustration of three different types of speech signal representation.

4 Feature reduction

Feature reduction process aims to reduce the dimensionality of the feature space, i.e., to reduce the number of features before classification. The feature reduction is performed for two reasons

1. **To remove redundant features**: Some features have strong mutual correlation and thus they don't provide any additional information. In this case only one representative feature of mutually correlated features is used instead of the whole group.
2. **To select most relevant features:** only best features in terms of discriminative power between considered classes are selected for further usage

Method for feature reduction can be divided into two main groups (Figure 4.1).

A. Methods for feature selection: the result of such methods is a subset of features selected from the original group based on a certain criterion. There are several criteria that might be employed for feature selection, such as the inter-class and intra-class distances, mutual information and the classification accuracy. The feature values remain unchanged after the feature selection process.

B. Methods for feature Transformation: These methods apply a mathematical transform function to the input feature space, which results in a new space of smaller dimensionality. The aim of such methods is to decorrelate the original features. Principal component analysis (PCA) is one of the most popular techniques for feature decorrelation.

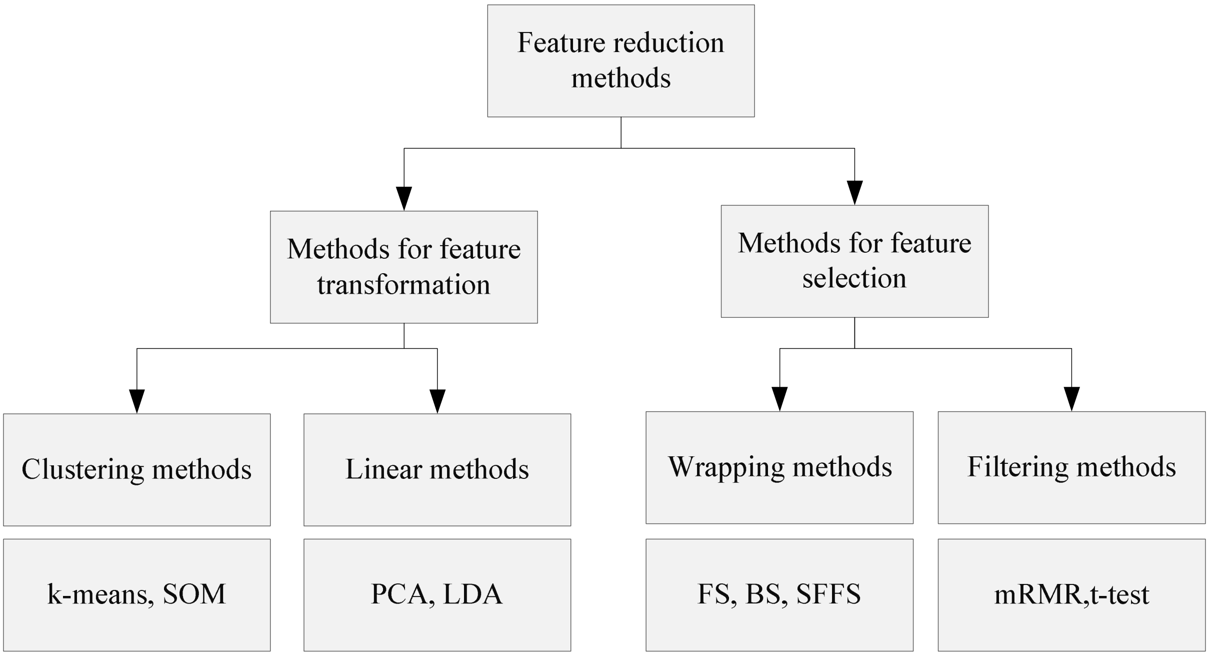


Figure 4.1: Categories of feature reduction techniques.

4.1 Principal component analysis

This is a very-well known method for reducing the number of dimensions of the input feature space based on the calculation of second order statistical moments. PCA outputs are numerical values known as principal components which are always uncorrelated, but may not be statistically independent. Moreover, these components are usually sorted in descending order according to their information content. PCA is successfully employed for emotion recognition from both speech and image modalities. It is often combined with classifiers which are extremely sensitive to correlated features such as *k*-NN. Another meaningful combination is to use PCA and GMM classifier with diagonal covariance matrix as this classifier doesn't effectively capture the correlation across features and thus, it is better to decorrelate the input feature space before classification.

4.2 *k*-means

This algorithm is used to partition the input feature space into clusters in which each feature vector belongs to the cluster with the nearest centroid (mean vector) [DHD12]. These centroids are then used to represent the whole original feature space.

4.3 Minimal-redundancy maximal-relevance criterion

This feature selection method proposed in [PLD05] aims to identify most discriminant features according to two criteria

1. Maximal relevance:

$$\max R(\mathrm{Z}, c), \quad R = \frac{1}{|\mathrm{Z}|} \sum_{x_i \in Z} I(f_i; c), \tag{4.1}$$

2. Minimal redundancy

$$\min O(\mathrm{Z}), \quad O = \frac{1}{|\mathrm{Z}|^2} \sum_{f_i, f_j \in Z} I\left(f_i; f_j\right), \tag{4.2}$$

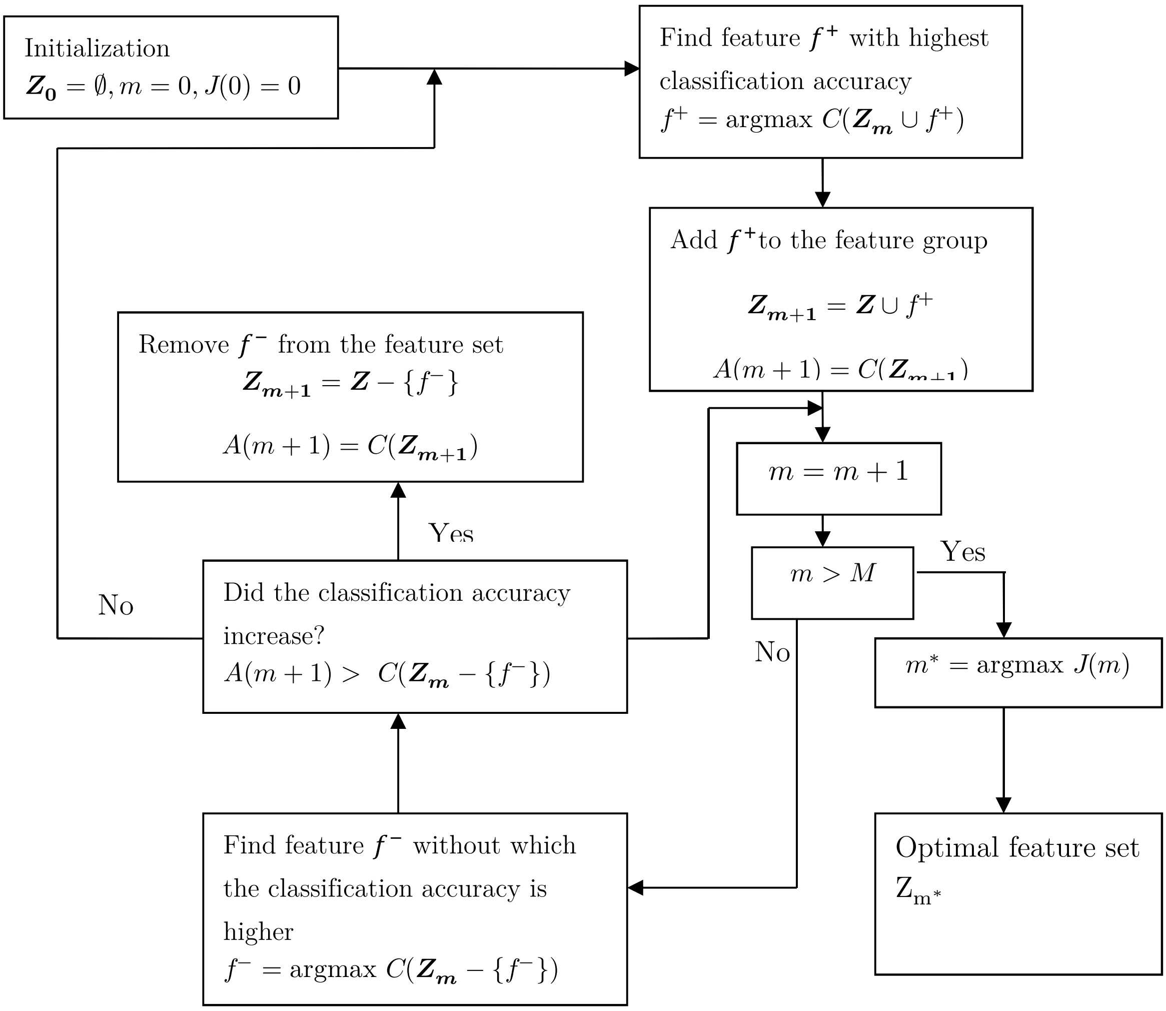
Where Z is the group of features under examination, $I(f_i; c)$ is the mutual information between feature f_i and class c and $I\left(f_i; f_j\right)$ is the mutual information between features. The criterion combining constraints (4.1) and (4.2) is called Minimal-redundancy maximal-relevance (mRMR). The advantages of mRMR algorithm are low computational complexity, classifier-independency and very effective selection of uncorrelated features.

4.4 Wrapping methods

These methods involve a classifier in their selection scheme and consider the classification accuracy as the only criterion for feature selection. Several methods based on wrapping have been proposed so far in literature. The simplest algorithm among them is Forward selection (FS). This algorithm starts with empty set of features and iteraltivley adds one feature in each iteration based on the classification accuracy. Once a feature is added it cannot be removed later. The disadvantage of forward selection is that it may fail to select features that are redundant. On the other hand, it may identify the subset of optimal features rapidly [WF05].

Backward Selection (BS) [LM98] starts with all features and iteratively removes one feature in each iteration. Again the criteria of removing features is the classfiication accuracy of an arbitaary classifier used within the algorithm. Comparing to forward selection, backward selection can effectively handle the redundant features. However, more iterations are needed to find the optimal feautre subset.

Sequential Floating Forward Selection (SFFS) is one of the most effective wrapping methods for feature selection propsed so far [Pud+94]. This method is similar to FS as it also works in an iterative manner and starts with empty set of features. However, the features selected after each iteration are removed one by one. If the removal of any feature results in increasing the classification accuracy, then the corresponding feature is permanently discarded from the feature set. This approach gurantees that the final set doesn't contain correlated features. The block scheme of SFFS algorithm is illustrated in Figure 4.2.


Initialization
$\boldsymbol{Z_0} = \emptyset, m = 0, J(0) = 0$
Find feature $\boldsymbol{f^+}$ with highest classification accuracy
$f^+ = \text{argmax } C(\boldsymbol{Z_m} \cup f^+)$
Add $\boldsymbol{f^+}$to the feature group
$\boldsymbol{Z_{m+1}} = \boldsymbol{Z} \cup f^+$
$A(m+1) = C(\boldsymbol{Z_{m+1}})$
Remove $\boldsymbol{f^-}$ from the feature set
$\boldsymbol{Z_{m+1}} = \boldsymbol{Z} - \{f^-\}$
$A(m+1) = C(\boldsymbol{Z_{m+1}})$
$m = m + 1$
Yes
Yes
No
$m > M$
Did the classification accuracy increase?
$A(m+1) > \ C(\boldsymbol{Z_m} - \{f^-\})$
No
$m^* = \text{argmax } J(m)$
Find feature $\boldsymbol{f^-}$ without which the classification accuracy is higher
$f^- = \text{argmax } C(\boldsymbol{Z_m} - \{f^-\})$
Optimal feature set
$\mathrm{Z_{m^*}}$

$\boldsymbol{Z_m}$ is the set of selected features in the m-th iteration.

$C(\boldsymbol{Z})$ is the classification function for feature set $\boldsymbol{Z}$.

$A(m)$ is the classification accuracy in m-th iteration .

M is the number of iterations.

Figure 4.2: Block scheme of SFFS algorithm.

5 Classification

Classification is the task of recognizing to which of a set of classes (categories) a new pattern belongs. This is performed by using a training set of instances whose category is previously known. The training instances are defined by various features. The mathematical algorithm that performs classification is known as a classifier.

5.1 Artificial neural network

Artificial Neural Networks (ANN) are biologically inspired pattern recognition classifiers consist of interconnected layers of neurons. The most known type of ANN is Feed Forward Back Propagation (FFBP). In the last decade, several more complex classification techniques based on ANN were have been proposed, such as Time Delay Neural Network and Echo state Network. All ANN classifiers have a big disadvantage which is the high computational complexity of training, especially when a big number of training instances or features are considered. More details on ANN can be found in [Yeg09].

5.2 *k*-nearest neighbor

This non-parametric classifier is simple, though very effective for both classification and regression. *k*-NN can achieve good classification accuracy when a small number of features are used. *k*-NN classifies each pattern to the corresponding class according to the number of nearest patterns in the feature space. The variable *k* defines the number of patterns that should be considered in the final decision.

5.3 Gaussian mixture models

GMM classifier represents a very effective tool for statistical modeling of the feature space using one or more Gaussian functions. Multivariate Gaussian distribution function is defined as

$$g(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\sum)}} \exp\left(-\frac{(F-\boldsymbol{\mu})^{\mathrm{T}} \sum{}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \tag{5.1}$$

Where

$\boldsymbol{x}$ is the input feature vector,

d is the number of dimensions (the same as the number of features),

$\boldsymbol{\mu}$ is the mean vector,
$\boldsymbol{\Sigma}$ is the covariance matrix defined as (5.2)

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{E}[(\mathrm{X}_1 - \mu_1)(\mathrm{X}_1 - \mu_1)] & \mathrm{E}[(\mathrm{X}_1 - \mu_1)(\mathrm{X}_2 - \mu_2)] & \dots & \mathrm{E}[(\mathrm{X}_1 - \mu_1)(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})] \\ \mathrm{E}[(\mathrm{X}_2 - \mu_2)(\mathrm{X}_1 - \mu_1)] & \mathrm{E}[(\mathrm{X}_2 - \mu_2)(\mathrm{X}_2 - \mu_2)] & \dots & \mathrm{E}[(\mathrm{X}_2 - \mu_2)(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})(\mathrm{X}_1 - \mu_1)] & \mathrm{E}[(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})(\mathrm{X}_2 - \mu_2)] & \dots & \mathrm{E}[(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})(\mathrm{X}_\mathrm{d} - \mu_\mathrm{d})] \end{bmatrix}. \qquad (5.2)$$

Gaussian mixture model comes from the linear combination of multiple Gaussian functions. This combination is mathematically defined as

$$G(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha_i \, g_i(\boldsymbol{x}), \qquad (5.3)$$

where
M is the number of Gaussian functions in the mixture,
$g(\boldsymbol{x})$ is the multivariate Gaussian distribution function,
α is the mixture weight.

It is obvious from (5.3) that the GMM classifier is defined by the individual parameters of Gaussian functions in the mixture, these parameters are the covariance matrix $\boldsymbol{\Sigma}$ and the mean vector $\boldsymbol{\mu}$. Moreover, each Gaussian function is weighted by parameter α. There exist several algorithms for estimation of GMM parameters, of which the most known is Expectation-Maximization (EM) algorithm. It is an iterative approach working in two steps 1) Expectation step to estimate the GMM parameters and set the likelihood function and 2) Maximization step where the GMM parameters are iteratively updated until the convergence of the likelihood function is detected.

A special case of GMM appears when only one Gaussian function with a diagonal covariance matrix is used. This classifier is known as Naïve Bayes Classifier (NBC) and since diagonal covariance matrix is employed in this classifier it doesn't capture the correlation between input features. However, the training of NBC is very simple comparing to GMM with full covariance matrix. Study in [RQD00] suggested that a combination of several NBC can successfully replace GMM with full covariance matrices for speaker verification.

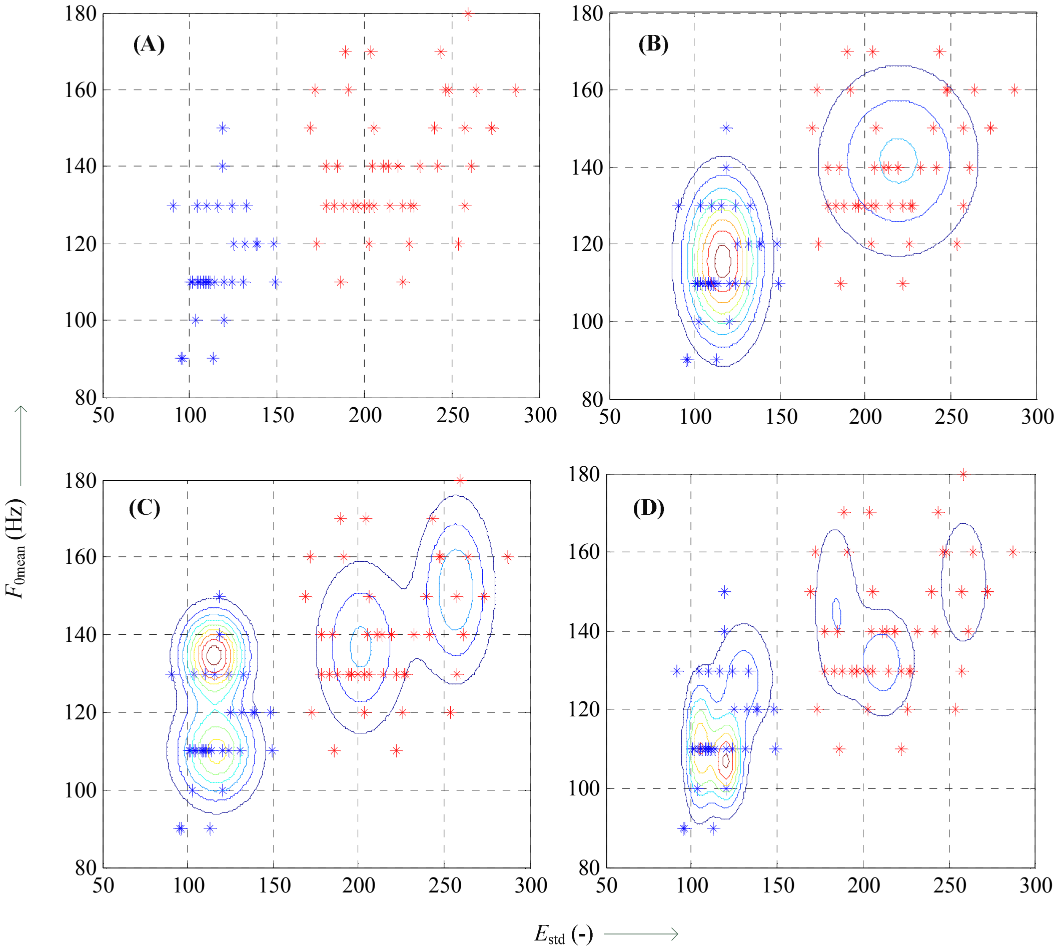


Figure 5.1: Illustration of GMM classifier. A: Feature space of two features (mean value of fundamental frequency $F_{0\text{mean}}$ and the standard deviation of temporal energy E_{std}) extracted from several utterances of two emotions (red: anger and blue: neutral) B: Feature space fitted by using one Gaussian function. C: Feature space fitted by using two Gaussian functions. D: Feature space fitted by using three Gaussian function.

5.4 Support Vector Machines

This classifier firstly proposed in [CV95] works on the basis of constructing a hyperplane in a high-dimensional space, SVM discrimination power is achieved by the hyperplane with the largest distance to the training data point of all classes considered. In other words, it maximizes the margin around the separating hyperplane. The decision function is defined by a subset of training data points known as support vectors. SVM classifiers are widely used in many pattern recognition domains because of their high classification accuracy and ability to work with high-dimensional feature vectors.

5.4.1 Linear SVM

The main concept of linear SVM is the inner (or dot) product between two vectors defined as $\mathrm{w}^{\mathrm{T}}\mathrm{x} = \sum_i w_i\, x_i$. Linear SVM is based on a linear discriminant function of the following form

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b, \tag{5.4}$$

vector $\mathbf{w}$ is the weight vector and b is called bias. The hyperplane that discriminate between features is defined as follows

$$\{\mathrm{x}\text{: } f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0\ \} \tag{5.5}$$

5.4.2 Nonlinear SVM

The idea of nonlinear SVM is to achieve linear separability by mapping the data points to a high-dimensional space. This goal is achieved by replacing the dot product in the linear classifier by a nonlinear kernel function. The generalized function of nonlinear SVM is given as

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathrm{x}) + b, \tag{5.6}$$

where $\boldsymbol{\phi}(\mathrm{x})$ is the kernel function.

There are several kernels proposed for nonlinear SVM classifiers, such as polynomial, hyperbolic tangent and radial basis. In own experiments, I employed SVM classifier with either linear or radial basis kernels.

5.5 Evaluation of classification performance

There are several approaches for the evaluation of classification performance. The most known among them is the confusion matrix. This matrix visualizes the classifier output in terms that each column represents the patterns in the classified class, while each row contains the patterns in the actual class.

The overall evaluation of classifier performance is usually delivered by two characteristics: the weighted accuracy and unweighted accuracy. These two characteristics are identical only when all testing classes have the same number of patterns.

The unweighted accuracy can be calculated as

$$A_{\mathrm{wa}} = \frac{100 N_{\mathrm{cor}}}{N_{\mathrm{p}}}, \tag{5.7}$$

where N_{cor} is the number of correctly classified patterns of all classes and N_{p} is the total number of patterns.

The weighted accuracy is given by

$$A_{\mathrm{uw}} = \frac{100}{C}\sum_{c=1}^{C} N_{\mathrm{cor}}^{\mathrm{c}}, \tag{5.8}$$

where $N_{\mathrm{cor}}^{\mathrm{c}}$ is the number of correctly classified patterns of class c and C is the number of classes.

A special case of classification appears when only two classes are considered, which is very typical in several domains, especially in healthcare, when the subjects are classified into two classes: positive and negative [Far+13]. As it will be shown later, some subtasks were treated as a binary classification task. The confusion matrix for binary classification is presented according to Table 5.1

Table 5.1: Confusion matrix for binary classification task.

TP True positive	***FP*** False positive
FN False negative	***TN*** True negative

From the confusion matrix in Table 5.1, the following evaluation characteristics can be derived

1. Precision

$$A_{\mathrm{pr}} = \frac{TP}{TP + FP} \tag{5.9}$$

2. Recall

$$A_{\mathrm{re}} = \frac{TP}{TP + FN} \tag{5.10}$$

3. F-Measure: The weighted harmonic mean of precision and recall

$$A_{\mathrm{fm}} = 2\frac{A_{\mathrm{pr}}A_{\mathrm{re}}}{A_{\mathrm{pr}} + A_{\mathrm{re}}} \tag{5.11}$$

4. Mathews correlation coefficient (MCC): This coefficient is a measure of binary classifier performance which takes into account all values of the confusion matrix reported in table 5.1. This indicator is regarded as a balanced measure which can be used even if the data distribution among the classes is not balanced. MCC output is a correlation coefficient with a value between −1 and +1.

$$A_{\mathrm{mcc}} = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5.12}$$

6 Emotion recognition from acted speech

6.1 Emotion recognition using BDES

The method described in this section was proposed in [AE08]. The aim was to find a new approach to improve the classification in comparison with the most recent reported work on BDES [LY07] in that time. The proposed approach works in two steps and employs features extracted on both segmental and suprasegmental level.

6.1.1 Database description

Berlin database of emotional speech (BDES) was developed at the Institute of Communication Science of the Technical University of Berlin [Bur+05]. The recordings were made in an anechoic room by 10 speakers (5 males and 5 females), who produced 10 German utterances in 7 different emotional states: angry, happy, feared, sad, disgusted, bored and neutral. Based on perception tests performed by 20 native listeners carried out to ensure the emotional quality and naturalness, 535 utterances with a recognition rate better than 80% and naturalness better than 60% were left out.

6.1.2 Feature extraction and selection

Four perceptual spectral features and their first and second differences were employed in this experiment. These features are MFCC, MELBS, PLP1 and PLP2. These features were extracted on segmental level from frames 250ms long with a 50% of overlap. This choice takes into account the results of several trial and error processes made evaluating the classifier performance on frame lengths ranging from 20 to 500ms, where the best classification rate was obtained for a frame length of 250ms.

6.1.3 Classifier selection

- GMM with one Gaussian function and a diagonal matrix for each state (emotion).
- FFBP-ANN with three layers, on input layer with 30 neurons, one hidden layer with 30 neurons and one output layer with 6 neurons. According to [Hua+01] this architecture meets the minimum requirements for the classification task under examination.
- *k*-NN, The number of nearest neighbours (*k*) was set to 5 based on trial-error process.

By using segmental features reported in the previous section, the GMM classifier outperformed the other classifiers and hence it was selected for further experiments. The results achieved for different classifiers and feature extraction schemes are illustrated in Figure 6.1.

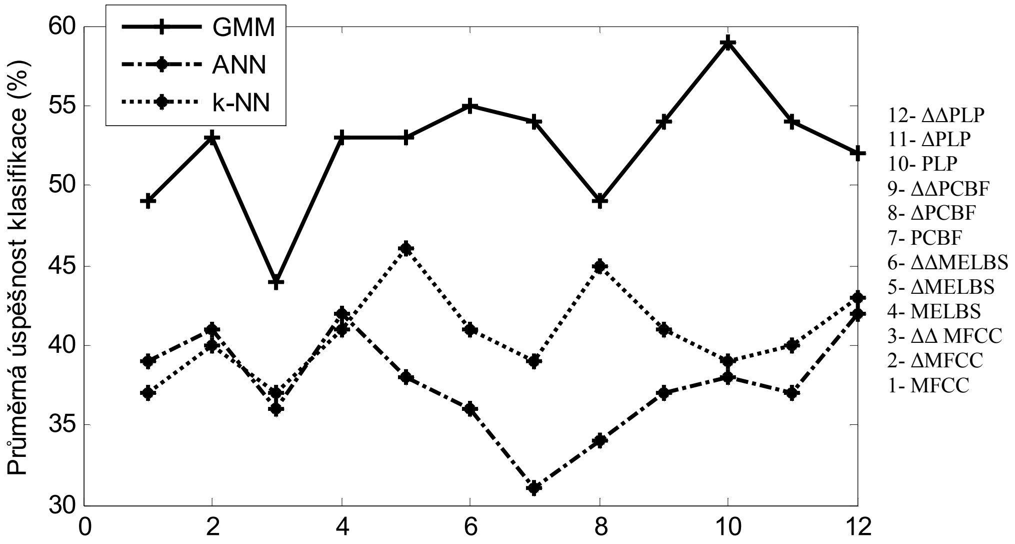


Figure 6.1: Classification accuracy for different classifiers and feature extraction schemes for BDES database.

Figure 6.2 displays the mean classification rates (averaged over the six emotions under examination) obtained using the encoding procedures and the GMM classifier discussed above (bars 1 up to 13). The best result (63%) was obtained through a combination of PLP, ΔPLP, PCBF and ΔΔMELB coefficients. The features included in this combination were identified using the Sequential Floating Forward Selection (SFFS) algorithm.

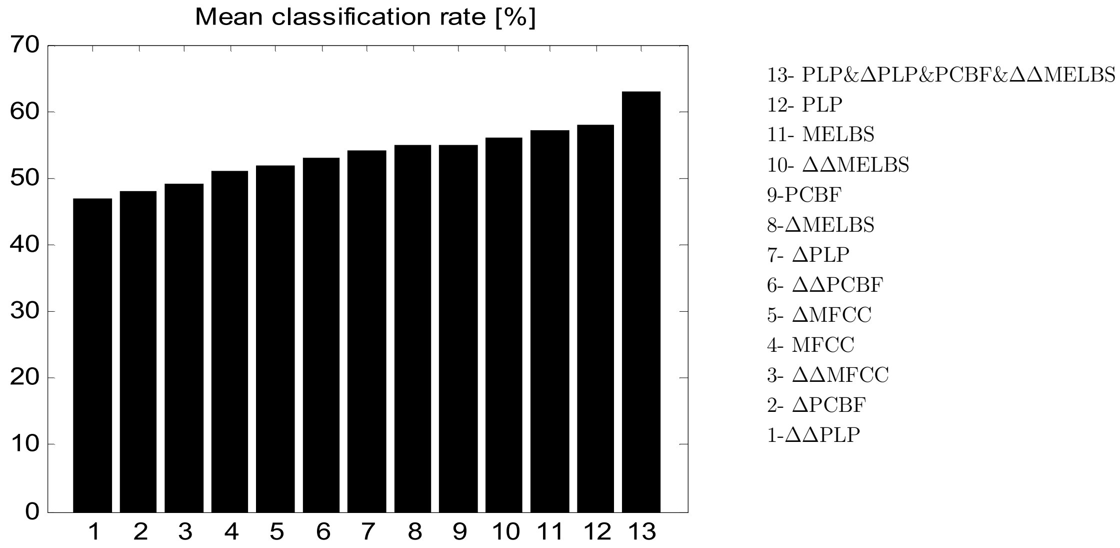


Figure 6.2: Mean classification rates for different perceptual features.

The details of the obtained results are displayed by the confusion matrix in Table 6.1. It can be noticed that satisfying classification rates are obtained only for anger (92%) and sadness (76%) emotional states. It can also be noticed that happiness emotional state is highly confused with an anger one, as well as a bored emotional state is confused with a neutral one. The high confusion between specific couple of emotional states suggests the need for a further processing and classifying step which should provide a way to overcome such problems. Moreover, since the GMM output is a likelihood valued, it should make sense to take into account couples of emotional states that have obtained from the first classifying step the highest likelihoods.

Table 6.1: Confusion matrix obtained within the first step (average classification rate: 63%)

	Anger	Boredom	Fear	Happiness	Sadness	Neutral
Anger	**76**	0	12	12	0	0
Boredom	0	**52**	4	0	4	40
Fear	8	4	**64**	20	0	4
Happiness	46	0	0	**54**	0	0
Sadness	0	4	0	0	**92**	4
Neutral	0	52	4	0	4	**40**

At the light of the above considerations, the second step considers only the two emotions that obtained the highest likelihoods scores within the first step in order to discriminate among them. The idea of choosing two emotions from the first step and not only one can be explained on a simple example: suppose that Bob (the first step of the approach) asked to guess the winner of a tournament which includes for instance six teams. Bob believe that there are two favorites which can win the competition and these two teams are almost equally strong; thus he will have a problem to determine which team will win the tournament, but he is almost sure that the winner will be one of these two strong teams. In this case he can ask someone (the second step of the approach) who better knows these teams to decide the winner. In the case at the hand, if it is assumed that the input emotion is classified correctly if it has the highest likelihood or second highest likelihood score, then, the mean classification rate will be about 94%.

6.1.4 Feature extraction within the second step

In the second step new feature extraction techniques were applied to the emotional speech utterances in order to identify for each couple of emotions a unique set of acoustic emotional features capable of discriminating between them. To this aim, prosodic and voice quality features including pitch, energy, zero crossing rate and harmonicity. These features were extracted from speech frames 125 ms long with a 50% of overlap. In this further processing step, the frame length was reduced in order to obtain acoustic vectors described by more prosodic and voice quality details. A total of 72 high-level features were obtained. Those that revealed to be relevant in the discrimination of the 15 couples of emotional states under examination are listed in Table 6.3. For each couple the best set of discriminate features was identified through the SFFS algorithm. The cross-emotion classification rates between each couple are reported in Table 6.2.

Table 6.2: Cross-emotion recognition within second classification step (average classification rate: 95.7 %).

	Anger	**Boredom**	**Fear**	**Happiness**	**Sadness**	**Neutral**
Anger	---	100	96	96	100	100
Boredom	100	---	96	100	96	90
Fear	88	92	---	88	92	96
Happiness	80	100	96	---	100	100
Sadness	100	96	96	100	---	92
Neutral	100	88	96	100	96	---

Table 6.3: List of selected high-level acoustic features for each couple of emotional states under examination using the Sequential Floating Forward Selection (SFFS) algorithm.

anger/boredom	pitch mean, position of pitch maximum, position of pitch first derivative maximum, pitch second derivative slope
anger/fear	pitch maximum, harmonicity mean, energy relative minimum
anger/happiness	energy first derivative relative maximum, position of energy first derivative minimum, harmonicity standard deviation, energy relative minimum, position of

	energy maximum, pitch first derivative minimum, pitch first derivative range, position of zero crossing ratio minimum
anger/sadness	pitch mean, pitch standard deviation
anger/neutral	pitch mean, pitch maximum, position of pitch first derivative minimum
boredom/fear	zero crossing ratio mean, pitch standard deviation, position of energy second derivative minimum, pitch relative minimum
boredom/happiness	pitch mean, pitch first derivative slope
boredom/sadness	pitch mean, relative maximum of pitch first derivative, relative minimum of pitch first derivative, zero crossing ratio maximum
boredom/neutral	pitch relative maximum, pitch first derivative relative maximum, position of pitch second derivative minimum
fear/happiness	pitch range, position of pitch first derivative maximum, zero crossing ratio relative minimum
fear/neutral	zero crossing ratio mean, harmonicity mean, pitch standard deviation, position of pitch first derivative minimum, pitch mean, energy relative minimum
fear/sadness	pitch relative minimum, zero crossing ratio relative maximum
happiness/sadness	pitch mean, pitch maximum, pitch standard deviation
happiness/neutral	pitch standard deviation, energy relative minimum, pitch mean
sadness/neutral	pitch mean, pitch standard deviation, energy standard deviation

Combining the two processing steps by considering the couple of emotions with the highest likelihoods in the first step as an input of the second step, the final confusion matrix obtained is shown in Table 6.4. Figure 6.3 illustrates the mechanism of the introduced approach on an example. It is worth to note that, even though at the output of the first step, the neutral emotional state obtained a highest likelihood score, boredom (which was the emotional state in input), was correctly classified as output of the second step.

Table 6.4: The final confusion matrix for BDES (average classification rate: 80.7%).

	Anger	Boredom	Fear	Happiness	Sadness	Neutral
Anger	**80**	0	8	12	0	0
Boredom	0	**80**	4	8	0	8
Fear	8	0	**76**	12	0	4
Happiness	24	0	0	**76**	0	0
Sadness	0	4	0	0	**92**	4
Neutral	0	12	4	0	4	**80**

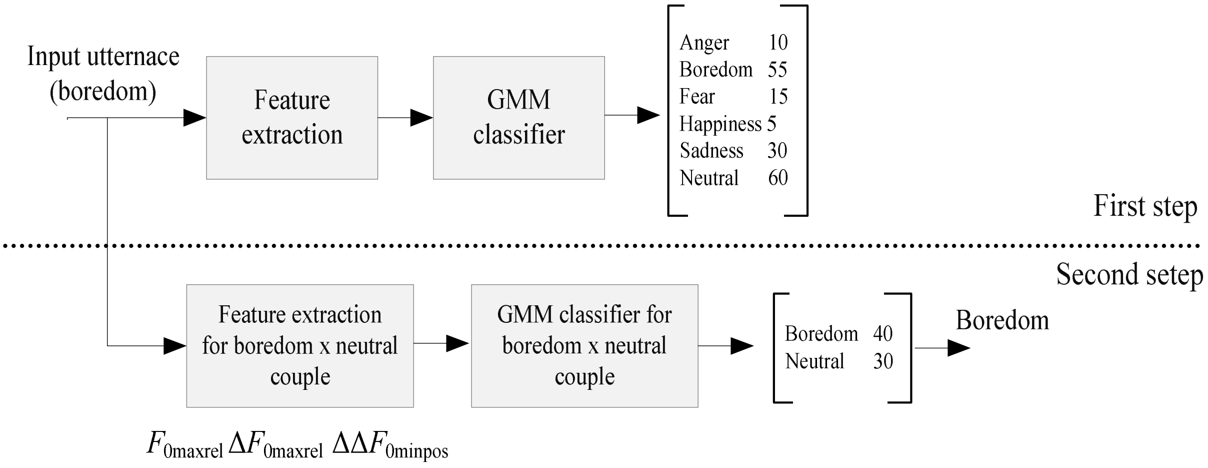

Input utternace (boredom)
Feature extraction
GMM classifier
Anger 10
Boredom 55
Fear 15
Happiness 5
Sadness 30
Neutral 60
First step
Second setep
Feature extraction for boredom x neutral couple
GMM classifier for boredom x neutral couple
Boredom 40
Neutral 30
Boredom
$F_{0\text{maxrel}}\,\Delta F_{0\text{maxrel}}\,\Delta\Delta F_{0\text{minpos}}$

Figure 6.3: An example of automatic vocal emotion recognition performed by "six by two approach".

6.1.5 Summary

This section proposed a new approach for automatic speaker-independent vocal emotion recognition. The main idea is in the splitting of the recognition task into two steps. The first step implements a coarse encoding and classification of the six emotional states under examination in order to identify among them the couple with the highest likelihoods. For this step, segmental-based representation of features were used and the features extracted were combined in order to obtain acoustic vectors best descriptive of emotional states. The second steps re-encode, through a set of prosodic and voice quality suprasegmental features, the emotional states that obtained the highest likelihood scores in the first step. The average classification rate obtained (80.7 %) by the proposed approach was better than the result (74.5%) reported by the most recent work of Lugger &Yang on speaker-independent recognition of emotional vocal expressions. It should be mentioned that the task of finding the best set of high-level features for separating within each couple of emotions is a very challenging issue. The features listed in Table 6.3 were obtained by applying the SFFS algorithm to the utterances in the Berlin database of emotional states. It is expected to obtain a different set of features with a different database

6.2 Emotion recognition using CDES

6.2.1 Database description

COAST 2102 database of Emotional Speech (CDES) [ERM09] contains data based on extracts from Italian movies whose protagonists were carefully chosen among actors and actresses that are largely acknowledged by the critique and considered capable of giving some very real and careful interpretations. The database consists of audio stimuli representing 6 basic emotional states: happiness, sarcasm/irony, fear, anger, surprise, and sadness [Ekm92].

For each of the above listed emotional states, 36 stimuli were identified, 18 expressed by an actor and 18 expressed by an actress, for a total of 216 audio stimuli. The stimuli were selected short in duration (the average stimulus' length was 2.5s, SD = ± 1s). This was due to two reasons: 1) longer stimuli may produce overlapping of emotional states and confuse the subject's perception; 2) emotional states for definition cannot last more than a few seconds and then other emotional states or moods take place in the interaction [OKJ06].

The emotional labels assigned to the stimuli were given first by two expert judges and then by three native judges independently. The expert judges made a decision on the stimuli carefully exploiting emotional information on facial and vocal expressions such as frame by frame analysis of changes in facial muscles, and F_0 contour, rising and falling of intonation contour, etc, as reported by several authors in literature.

6.2.2 Feature extraction

As it was mentioned in chapter 5, features usually used to recognize emotional vocal expressions can be divided into three main groups: prosodic features, voice quality features and spectral features. In the present approach the prosodic and voice quality features were extracted on suprasegmental level whereas the spectral features were extracted on segments level. The following prosodic and voice quality features were considered: Fundamental frequency, temporal energy, Formants frequencies, formants bandwidths and harmonicity. Beside the original waveforms of mentioned features, their first and second differences were also taken into account. In summary a total number of 225 sprasegmental (high-level) including mean, maximum, minimum, slope, standard deviation and other statistical measurements measured in speech frames 125 ms long with 50% of overlap were considered. Regarding segmental features, the following perceptual spectral features were extracted from each frame: MFCC, PLP4, and MELBS.

6.2.3 Feature reduction

Feature extraction could be carried out either on frame (segmental) or utterance (suprasegmental) level. On the frame level, each frame is considered as a single input training pattern. The classification (decision making) process is then applied on each frame separately. The final decision about the utterance emotion is taken then for instance, according to the appearance of each class (emotion) in the obtained result.

If each utterance is considered as a single input pattern in the classification process, it should be taken into account the different number of segments obtained from each utterance, that yields feature vectors with various lengths after the concatenation of the spectral characteristics (such as MFCC) extracted from the frames of each utterance under examination. A possible solution to this inconvenient is to zero padding in order to have vectors of the same length. However, this solution didn't give satisfying results in terms of classification accuracy. A better way to handle the problem would be to use feature space reduction techniques such as Vector Quantization algorithms like *k*-means or Principal Component Analysis (PCA). In the present approach a relatively simple space reduction method is proposed based on spectral vector averaging that had provided better results than PCA or *k*-means on the data under examination. The principle of the proposed space reduction approach called Temporal Mean Vector (TMV) is shown in Figure 6.4 and is applied according to the following steps:

1. The spectral feature vectors are extracted from speech frames 250 ms long with 50% of overlap. The frame length was set through several trial and error processes. It could be surprising that the chosen length (250ms) is significantly longer than that usually used for phoneme-based speech applications as observed by Apolloni et al. [AAE00] "*affective acoustic parameters are characterized by a lower rate of variability in time than linguistic ones*".
2. The spectral feature vectors obtained from the first, second and third part of a given utterance are separately averaged, i.e., the centroids of corresponding spectral feature vectors are computed. All three utterance parts have the same length. The purpose of dividing utterances into parts was to include temporal information in the final feature vector. The centroids are computed as follows:

$$\mathbf{x}_m^j = \frac{\sum_{n=\alpha_1}^{\alpha_2} \mathbf{x}_n^j}{\left\lfloor \frac{N}{3} \right\rfloor}, \qquad j = 1, 2, 3. \tag{6.1}$$

Where $\mathrm{x}_{\mathrm{m}}^{\mathrm{j}}$ is the *j*-th mean feature vector (centroid) of *j*-th utterance's part, *N* is the number of extracted spectral feature vectors, $\alpha_1 = \lfloor \frac{(j-1)N}{3} \rfloor$and $\alpha_2 = \lfloor \frac{jN}{3} \rfloor - 1$.

The final spectral feature vector x_{sp} is obtained by concatenating the three centroids:

$$\mathbf{x}_{\mathrm{sp}} = [\mathbf{x}_m^1, \mathbf{x}_m^2, \mathbf{x}_m^3]^{\mathbf{T}} \tag{6.2}$$

For prosodic and voice quality features, the SFFS algorithm was exploited in order to identify features that showed the maximum capability of discriminating within couples of emotions (we shall refer to this as emotion coupling). The SFFS algorithm was applied separately on the prosodic-voice quality features and the spectral features.

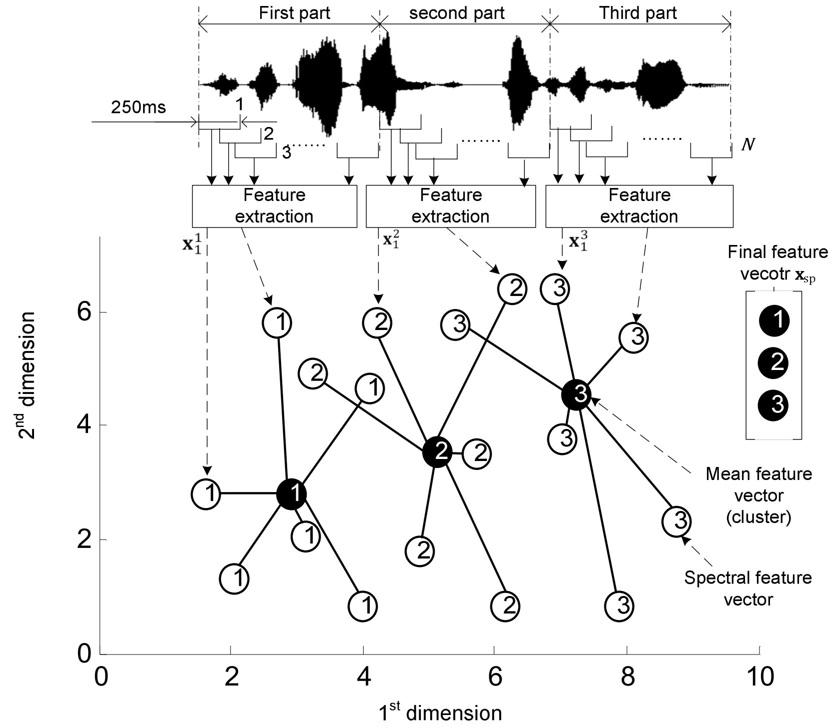


Figure 6.4: The principle of Temporal Mean Vector Method for feature reduction.

The advantage of the SFFS algorithm is that it identifies the best features according to their classification accuracy by using an arbitrary classifier. In own experiments, a GMM classifier (with one Gaussian per class) was used both for feature selection via the SFFS algorithm and for the overall validation of the proposed vocal emotion recognition algorithm. Thus, it could be stated that the features were chosen with a high level of reliability since each GMM classifier used in the final proposed system has already found its optimal features through the SFFS algorithm. The features that showed the best performance in distinguishing within couples of emotions are reported in Table 6.5.

Table 6.5: Selected prosodic, voice quality and spectral features for all couples of emotional states using the SFFS algorithm.

Emotions couples	**prosodic and voice quality features**	**Spectral features**
Anger X fear	energy slope minimum, pitch slope mean, range of pitch second derivative, mean of energy second derivative	ΔMFCC, ΔMELB
Anger X happiness	energy range, second formant mean	MELB, ΔPLP
Anger X irony	mean of pitch second derivative, pitch minimum, pitch standard deviation, second formant bandwidth standard deviation	PLP, ΔPLP

Anger X sadness	pitch range, maximum of energy second derivative, pitch mean, energy median, second formant standard deviation	PLP, MELBS
Anger X surprise	maximum of pitch second derivative, maximum of energy second derivative, pitch slope standard deviation, mean of pitch second derivative, first formant mean, harmonicity mean.	ΔMELBS, ΔPLP
Happiness X fear	jitter, energy slope maximum, harmonicity mean	MFCC,ΔPLP,ΔMELBS
Happiness X irony	Jitter, pitch minimum, mean of pitch second derivative, maximum of pitch first derivative, maximum of pitch second derivative	ΔPLP,MELBS
Happiness X sadness	pitch relative maximum, range of energy second derivative, harmonicity standard deviation	MELBS,ΔMELBS
Happiness X surprise	jitter, standard deviation of energy second derivative, second formant standard deviation, harmonicity maximum	ΔΔPLP,ΔΔMELBS
Irony X fear	mean of energy second derivative, pitch slope maximum, pitch slope minimum, pitch standard deviation, mean of pitch first derivative, jitter, standard deviation of energy second derivative	MFCC,ΔPLP,ΔΔPLP
Sadness X fear	energy slope minimum, pitch median, First formant frequency mean, first formant bandwidth mean	MFCC,PLP,ΔPLP, ΔΔMELBS
Sadness X irony	energy median, Energy slope minimum, energy range	MEBLS, MFCC, ΔPLP
Sadness X surprise	energy slope maximum, pitch relative maximum, minimum of pitch second derivative	PLP,MELBS,ΔMFCC, ΔΔMFCC
Surprise X fear	range of pitch second derivative	MFCC , PLP, ΔPLP
Surprise X irony	mean of pitch second derivative, energy slope standard deviation, relative maximum of energy first derivative, energy slope standard deviation, range of pitch first derivative	ΔPLP, ΔΔPLP, MELBS

6.2.4 Classification

The proposed classifier uses two classification techniques fused together in two classification steps as illustrated in Figure 6.6.

The first step (Figure 6.5) is used to train each sub-classifier $D^{(i)}$ to distinguish within couples of emotions. The likelihoods output by each classifier are then multiplied by each other. The final decision about the classes scores $\aleph_{\omega}$, $\aleph_{\omega'}$ is made according to the following formula

$$\aleph_{\omega} = \begin{cases} 1 & \text{for} \quad P(\omega|\mathbf{x}_{\text{sp}})P(\omega|\mathbf{x}_{\text{pv}}) > P(\omega'|\mathbf{x}_{\text{sp}})P(\omega'|\mathbf{x}_{\text{pv}}) \\ 0 & \text{for} \quad P(\omega|\mathbf{x}_{\text{sp}})P(\omega|\mathbf{x}_{\text{pv}}) < P(\omega'|\mathbf{x}_{\text{sp}})P(\omega'|\mathbf{x}_{\text{pv}}) \end{cases}, \tag{6.3}$$

where P(ω|x) is the posterior density distribution of the emotion category ω and $\mathbf{x}_{\text{sp}}$, $\mathbf{x}_{\text{pv}}$ are spectral and prosodic-voice quality feature vectors respectively.

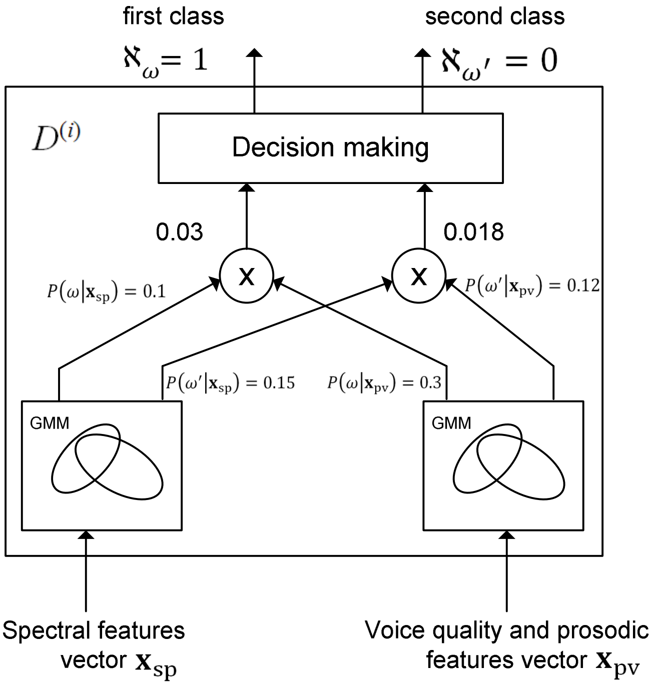

first class
second class
$\aleph_{\omega} = 1$
$\aleph_{\omega'} = 0$
$D^{(i)}$
Decision making
0.03
0.018
X
X
$P(\omega|\mathbf{x}_{\mathrm{sp}}) = 0.1$
$P(\omega'|\mathbf{x}_{\mathrm{pv}}) = 0.12$
$P(\omega'|\mathbf{x}_{\mathrm{sp}}) = 0.15$
$P(\omega|\mathbf{x}_{\mathrm{pv}}) = 0.3$
GMM
GMM
Spectral features vector $\mathbf{x}_{\mathrm{sp}}$
Voice quality and prosodic features vector $\mathbf{x}_{\mathrm{pv}}$

Figure 6.5*:* The fusion of prosodic-voice quality and spectral feature vectors.

The GMM parameters (mean vector and covariance matrix) were estimated using Estimation-Maximization (EM) algorithm initialized using the *k*-means clustering algorithm. Only one Gaussian was used to model each emotion category. This applies to all GMM used in the classification scheme.

The second step uses a simple perceptron with 6 neurons (one for each emotion).

The neurons have a linear transfer function described by

$$y_{\omega} = \sum_{i=1}^{N} \aleph_i^{\omega}\,, N = 6. \tag{6.4}$$

6.2.5 Results

The proposed algorithm was validated using the leave-one-speaker-out validation technique. The classification rates within couples of emotions are reported in Table 6.6. The final confusion matrix is shown in Table 6.7 and the average classification rate was 60.7%.

Figure 6.7 illustrates the differences in accuracy between the automatic and human subjective classification. The correlation between them is high (the normalized correlation here is R=0.79) even though the proposed system performs better for all the emotions except irony.

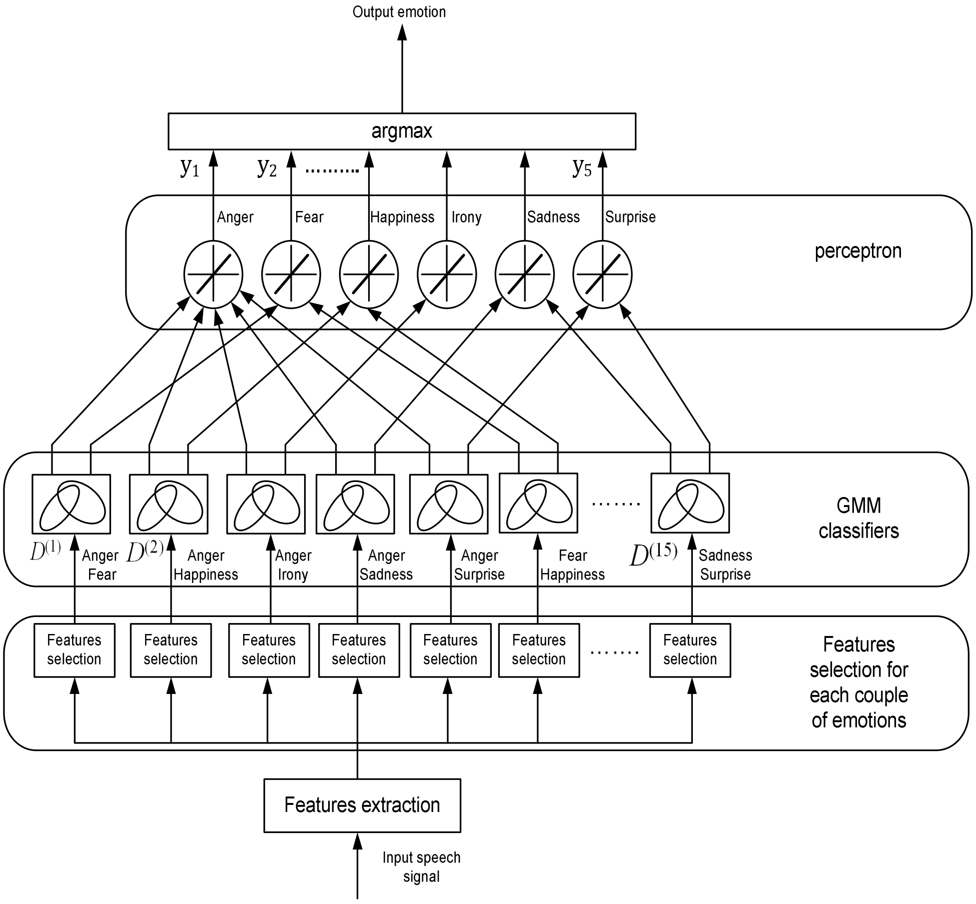


Figure 6.6: Proposed classifier based on emotion coupling.

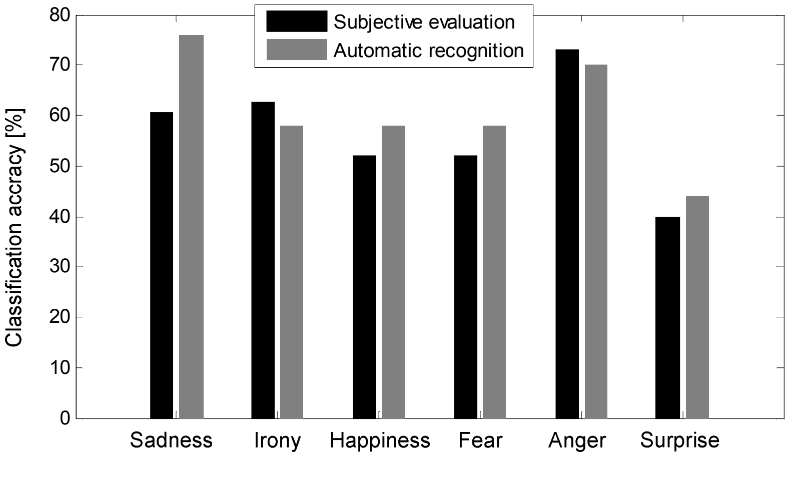


Figure 6.7: Classification performance achieved by the proposed system and by human subjects.

Table 6.6: Cross-emotion recognition within couples of emotions (average classification rate: 76.4 %).

	Anger	Fear	Happiness	Irony	Sadness	Surprise	average
Anger	-	73	70	76	84	77	76
Fear	71	-	72	87	76	68	75
Happiness	73	74	-	76	73	72	74
Irony	74	85	78	-	79	84	80
Sadness	89	80	77	81	-	85	83
Surprise	71	62	66	76	73	-	70

Table 6.7: The final confusion matrix for the six emotions under examination (average classification rate: 60.7 %).

	Anger	Fear	Happiness	Irony	Sadness	Surprise
Anger	**70**	6	6	12	6	0
Fear	9	**58**	9	9	12	3
Happiness	12	6	**58**	3	15	6
Irony	12	0	15	**58**	9	6
Sadness	0	6	12	3	**76**	3
Surprise	12	14	12	3	15	**44**

6.2.6 Summary

In this section, a new approach was proposed for automatic speaker-independent vocal emotion recognition validated by using the COST 2102 Italian database of emotional speech. The proposed system is mainly based on a new classifier consisting of the fusion of a simple perceptron and fifteen GMM classifiers designed to distinguish within couples of emotions under examination. The optimal spectral, prosodic and voice quality features that showed high discriminative power were chosen by using the SFFS algorithm.

The mean classification rate of the presented system is 60.7% with a significant improvement of 20.7% with respect to the baseline result (40%) obtained with an automatic system previously proposed in section that had provided the best classification results on the BDES database used as benchmark in recent literature. The obtained results are slightly better than those achieved by human subjects (56.5% on average) even though there is a high correlation ($R = 0.79$) among them. In the light of above results, it is difficult to answer why the presented system performed better than human subjects, since the difference in terms of classification accuracy among them is not significant. The results reported in this section were published in [Ata+09].

6.3 Advanced analysis of high-level features

In previous sections it was proven that suprasegmental (high-level) features perform very well in terms of emotion classification from speech. However, little attention has been paid so far to the statistical analysis of these features. The next section is hence devoted to emotion recognition by using only high-level features by exploiting two different databases of acted speech: BDES and CDES. The results of this section was published in [AES11]

6.3.1 Feature extraction

The list of extracted features for this experiment is reported in Table 6.8 and the feature extraction process is illustrated in Figure 6.8. In the first step, the speech signal is segmented by using a 32ms Hamming window with 50% overlap. The following Step involves the computation of features considered from each speech segment. In the third step, the high-level features are extracted from the segmental features and finally, these features are concatenated into the final feature vector used for classifier training. Beside the features listed in table 6.8, the first and second differences (Δ, ΔΔ) of these features were also considered

One can see from Figure 6.8 that the segmental feature vectors obtained from the first, second and third part of a given utterance are separately processed. All three utterance parts have the same length. The purpose of dividing utterances into parts was to include temporal information in the final feature vector. I also considered the case when no temporal information is involved. The results for both approaches are reported in section 6.3.3.

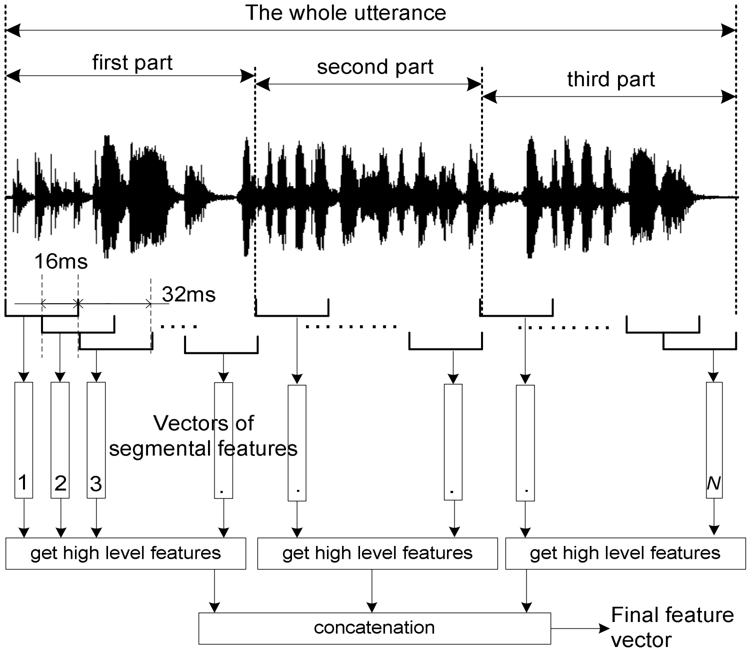


Figure 6.8: Feature extraction process.

Table 6.8: List of considered features for the purpose of analyzing high-level characteristics.

Feature group	Abbreviation	No.
Perceptual spectral features		
Mel Freq. Cepstral Coefficients	MFCC	20
Human Factor Cepstral Coefficients	HFCC	20
Linear Frequency Cepstral Coefficients	LFCC	20
Perceptual Linear Predictive 1	PLP1	21
Perceptual Linear Predictive 2	PLP2	21
Perceptual Linear Predictive 3	PLP3	21
Perceptual Linear Predictive 4	PLP4	11
MEL Bank Spectral Coefficients	MELBS	20
Human Factor Bank Spectral Coefficients	HFBS	20
Linear Frequency Bank Spectral Coefficients	LFBS	20
LPC based features		
Linear Predictive Coefficients	LPC	10
Linear Predictive Cepstral Coefficients	LPCC	10
Adaptive Component Weighting	ACW	10
Wavelet based features		
Subband Based cepstral Coefficients	SBC	24
Wavelet Decomposition	WADE	8
Modulation energy based features		
Mel Spectral Modulation Energy	MSME	1
4Hz Modulation Energy	4HzME	1
Spectral Features	SF	5
Fundamental frequency	F0	1
Formant frequencies	Fx	5
Formant Bandwidths	Bx	5
Harmonicity	H	1
Temporal Energy	TE	1
Teager Energy Operator	TEO	1
Zero Crossing Ratio	ZCR	1
overall		**278**

The following high-level features are tested in this experiment: mean, median, standard deviation, maximum, minimum, range, relative range, relative maximum, relative minimum, position of maximum, position of minimum, relative position of maximum, relative position of minimum, slope, kurtosis, skewness, linear regression coefficient, linear regression error, Pearson's skewness coefficient, 3rd, 4th, 5th and 6th moment, 1%, 5%, 10%, 20%, 30%, 40%, 60%, 70%, 80%, 90%, 95% and 99% percentile.

6.3.2 Feature selection

The feature selection process involves two steps: In the first step, the minimum Redundancy Maximum Relevance (mRMR) algorithm was employed in order to reduce the feature number to 200. This algorithm shows a very good performance in comparison with similar techniques. Moreover, it is recommended to use mRMR as a preprocessing step before applying any wrapper method. The second feature extraction step is carried out using SFFS algorithm.

6.3.3 Classification

The classification of a feature vector ***F*** is done following the algorithm described in the pseudocode below. This algorithm, called emotion coupling, was introduced in section 6.2.

```
// Indices of all possible emotional pairs
P= {1,2; 1,3; 1,4; 1,5 ; 1,6 ; 2,3 ; 2,4 ; 2,5 ; 2,6 ; 3,4 ; 3,5; 3,6; 4,5; 4,6; 5,6}

// Score vector
v= {0,0,0,0,0}

// Incrimination of emotional couples
FOR $i$=1 TO 15
// $Z^i_{\text{opt}}$ is a vector of optimal features for the i-th couple of emotions
// $C^i$ is the GMM classifier trained for the i-th couple of emotions

    $y_{\text{out}} = C^i(\mathbf{F}(Z^i_{\text{opt}}))$

    $\mathbf{v}[\mathbf{P}[i, y_{\text{out}}]] = \mathbf{v}[\mathbf{P}[i, y_{\text{out}}]]+1$

ENDFOR

//The output emotion is that obtained the highest score
$em = \text{argmax}(\mathbf{v})$
```

The GMM parameters (mean vector and covariance matrix) were estimated using Estimation-Maximization (EM) algorithm initialized using the *k*-means clustering algorithm. Only one Gaussian was used to model each emotion category. The emotion couples for both databases are reported in Table 6.9.

Table 6.9: Emotion couples for BDES and CDES and their corresponding indices.

	BDES	CDES		BDES	CDES
1	Anger x boredom	Anger x fear	9	Boredom x sadness	fear x surprise
2	Anger x fear	Anger x happiness	10	Fear x happiness	Happiness x irony
3	Anger x happiness	Anger x irony	11	Fear x neutral	happiness x sadness
4	Anger x neutral	Anger x sadness	12	Fear x sadness	Happiness x surprise
5	Anger x sadness	Anger x surprise	13	Happiness x neutral	irony x sadness
6	Boredom x fear	Fear x happiness	14	Happiness x sadness	irony x surprise
7	Boredom x happiness	Fear x irony	15	Neutral x sadness	sadness x surprise
8	Boredom x neutral	Fear x sadness			

The classification results within the couples of emotions are shown in Tables 6.11 and 6.12. The average classification accuracies for either cases when the temporal information is or not considered are summarized in Table 6.10.

Table 6.10: Average classification accuracies. Y: Temporal information is considered, N: Temporal information is not considered.

BDES		CDES	
N	Y	N	Y
98.41%	98.73%	92.30%	96.36%

Table 6.11: Cross-emotion recognition results within couples of emotions for BDES. Y: Temporal information is considered, N: Temporal information is not considered.

	anger		boredom		fear		happiness		neutral		sadness	
	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
anger	-	-	100	100	96.8	96.1	87.4	92.3	100	100	100	100
bored	100	100	-	-	98	96.1	100	100	96.3	94.8	98	100
fear	95.9	95.6	97	100	-	-	95.5	97.1	100	100	100	100
Happ	96.9	96.3	100	100	96.9	98.9	-	-	100	100	97	100
neut	100	100	98.7	98.2	100	100	100	100	-	-	100	100
Sad	100	100	99	100	100	100	99	100	100	100	-	-

Table 6.12: Cross-emotion recognition results within couples of emotions for CDES. Y: Temporal information is considered, N: Temporal information is not considered.

	anger		**boredom**		**fear**		**happiness**		**neutral**		**sadness**	
	N	**Y**	**N**	**Y**	**N**	**Y**	**N**	**Y**	**N**	**Y**	**N**	**Y**
anger	-	-	100	100	97.1	91.2	76.5	94.2	97.1	100	94.1	97.1
bored	97.2	94.5	-	-	97.2	97.2	97.2	94.5	94.4	94.5	80.5	94.5
fear	94.6	86.1	91.9	97.3	-	-	94.6	94.6	81.1	91.9	94.6	100
Happ	88.6	91.4	94.3	97.1	100	97.1	-	-	97.1	94.2	97.1	97.1
neut	97.1	97.1	94.3	97.2	94.3	97.2	97.1	100	-	-	91.4	97.1
Sad	89.2	89.2	78.4	97.2	86.5	94.6	89.1	94.6	86.5	97.3	-	-

6.3.4 Results

This section reports the results of analysis made within the optimal feature group obtained by applying the procedure described in section 6.3.2.

Table 6.13: Number of selected high-level characteristics for each feature group. B: BDES. C: CDES.

Feature	**B**	**C**	**Feature**	**B**	**C**	**Feature**	**B**	**C**
MFCC	7	6	LPCC	0	0	F4	0	0
HFCC	12	8	ACW	1	2	F5	2	1
LFCC	2	3	WADE	3	18	B1	0	0
PLP1	1	2	SBC	0	2	B2	0	1
PLP2	7	7	4HzME	1	0	B3	0	4
PLP3	3	3	MSME	0	0	B4	0	0
PLP4	6	6	SF	5	1	B5	0	0
MELBS	2	6	F0	0	0	H	0	0
HFBS	1	3	F1	0	0	TE	1	1
LFBS	2	3	F2	0	0	TEO	0	0
LPC	2	7	F3	0	0	ZCR	0	1

Figure 6.9 illustrates the number of high-level characteristics extracted from the entire utterance and its parts. The number of high-level characteristics extracted from the original segmental feature waveforms and their first and second differences is illustrated as well.

It appears evident from Table 6 that the perceptual spectral features such as PLP, MFCC, MELBS and MFCC make the largest part among all features considered. Hence, it is worth to focus more on this feature set. The experiment was set as follows: the high-level characteristics extracted from the perceptual spectral features were chosen within the top 200 ranked features by mRMR algorithm. Subsequently, the filters from which the high-level characteristics were obtained were identified; this helped to make an idea about which frequency ranges are important from the emotion recognition point of view.

Figure 6.10 illustrates the histograms of frequencies in Hz obtained from the previous experiment. It also illustrates the probability density estimate based on normal kernel function

Table 6.14: The ten most frequent high-level characteristics among the selected optimal feature group.

	CDES	**No.**	**BDES**	**No.**
1	Minimum	24	Slope	14
2	Maximum	21	Minimum	9
3	Range	14	30% percentile	8
4	Position of maximum	11	5% percentile	8
5	Linear regression coefficient	9	Relative minimum	8
6	Skewness	9	20% percentile	7
7	90% percentile	9	1% percentile	7
8	20% percentile	8	40% percentile	7
9	Slope	7	70% percentile	6
10	Standard deviation	6	90% percentile	5

6.3.5 Summary

Based on the results achieved from our experiments, it can be concluded that

1. A bigger number of high-level characteristics were needed to achieve the highest possible classification accuracy for CDES in comparison with BDES. This can be explained as the emotions expressed in CDES are more realistic in comparison with BDES and those need more parameters to be distinguished.
2. The high-level characteristics extracted from the perceptual spectral features have the strongest distinguishing capability for both databases among other features. These features can replace with a high efficiency the classical prosodic and voice quality features such as pitch, temporal energy and formants.
3. The high-level features derived from the wavelet coefficients significantly contribute to the classification process for CDES database.

4. The Human Factor Cepstral Coefficients (HFCC) appeared to be more valuable than the classical Mel Frequency Cepstral Coefficients (MFCC) in terms of emotion recognition from speech.

5. The minimum value is the most valuable high-level feature for CDES and the second most valuable feature for BDES.

6. The percentiles are the most frequently reported high-level characteristics within the top ten ranked features for the BDES database.

7. The temporal information doesn't bring a significant improvement in terms of recognition rate in case of BDES. On the contrary, the inclusion of the temporal information increased the classification accuracy of CDES utterances by 4%.

8. The major part of optimal high-level characteristics was selected from the entire utterance for BDES. On the other hand, the utterance subparts appear to be a carrier of important information in case of CDES, especially the second part.

9. Similarly to the point mentioned above, the high-level characteristics extracted from the first and second segmental feature differences contribute more to the optimal feature group in case of CDES.

10. The probability density estimate in Figure 6.9 showed that the selected frequencies are mostly concentrated around the values of 250Hz and 850Hz for BDES and CDES respectively. This is a very interesting finding since the first central frequency can be explained as the most important high-level characteristics were extracted from filters located in the range of speech fundamental frequency. On the contrary, it is quite difficult to explain why the central frequency of CDES is much higher. One of the possible explanations can be that the formant frequency around 850Hz carry information about the emotional content

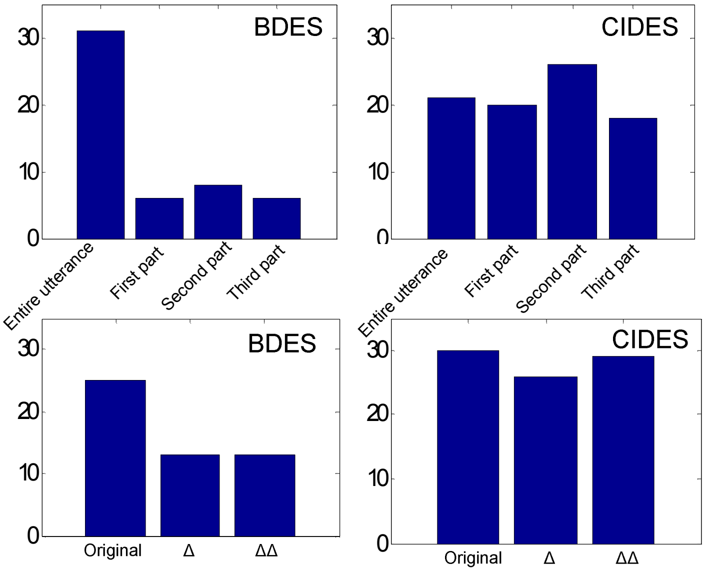


Figure 6.9: The number of high-level characteristics extracted from the entire utterance and its parts (up) and the number of high-level characteristics extracted from the original segmental feature patterns and their first and second differences (down).

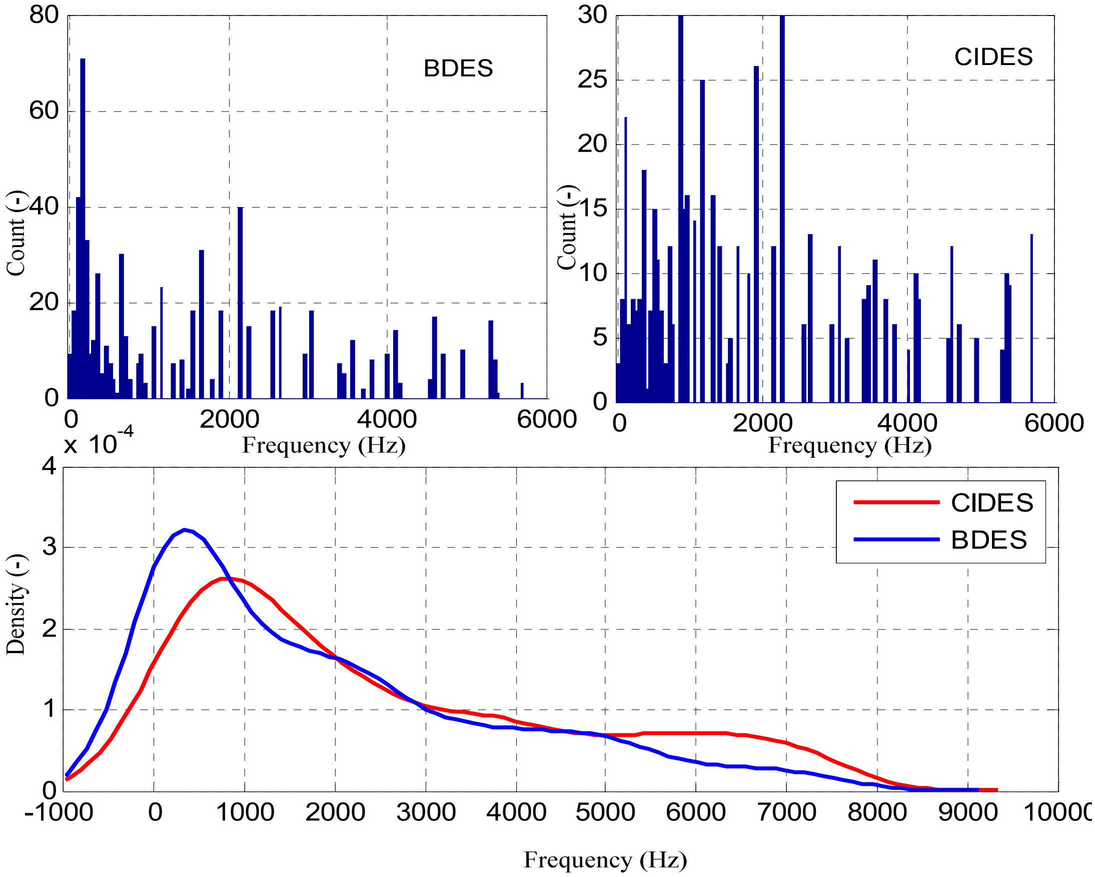


Figure 6.10: Histograms of frequencies in Hz and the probability density estimate.

7 Emotion recognition form spontaneous speech

This chapter deals with experiments made on spontaneous speech databases; the first section is devoted to the description of speech corpus acquired from European call centers, the second section proposes a method based on regression for emotion recognition from Slavic speech and third section presents a complex classification system for emotion recognition for all considered languages.

7.1 Spontaneous speech databases

This section presents the speech databases used in developing the system for emotion recognition from spontaneous speech. These databases were collected from different call centers in Europe in cooperation with Czech company Retia.

The collected data are based on extracts (short utterances) from dual-channel agent-client phone call records obtained from real call centers, mostly focusing on costumers' support services. The collaborating call centers which are based in nine countries: Czech Republic, Slovakia, Poland, Russia, Germany, England, Spain, Italy and France, provided us with raw speech records of corresponding languages. For each language, speech extracts were selected and subsequently labeled according to this format: *em_g_o_sp_la_abcdef*.wav, where *em* is the emotion identifier, *g* is the speaker's gender, *o* is the speaker's age, *sp* is the speaker's ID, *la* is the language ID and *abcdef* are the results of the subjective evaluation. For example, speech record "an_m_30_09_ru_7001002.wav" was labeled as anger; the speaker is Russian male aged 30 identified in the database by number 09. 7 listeners out of 10 agreed that the speaker expressed anger, 1 listener marked this utterance as neutral and 2 listeners marked it as other emotion.

All speech records are stored in PCM *.wav format with a sampling rate of 8 kHz and 16 bit quantization.

The construction of speech databases was carried out by the following procedure

1. Primary selection: After contacting several call centers in order to acquire the speech material, it was necessary to select appropriate speech records from these centers. Only records longer than three minutes were taken into account, this is due to the fact, that for shorter records it is not expected that the dialogue will be sufficiently developed and thus emotion expressions will be bland. Additionally, all records containing personal data of callers were discarded.

2. Subjective evaluation: Short utterance (2-4 seconds long in average) were extracted from the original phone calls and subsequently judged by native listeners in order to evaluate the emotional state of each utterance. The subjective evaluation tests were carried out by using GoEmotionally tool (Appendix B).

Table 7.1 gives an overview of spontaneous speech databases in terms of basic statistics and number of listeners (subjective evaluators). Tables 7.2 to 7.10 reports the results of subjective evaluation tests for each language

Table 7.1: Overview of spontaneous speech databases.

	Number of listeners		**Age of listeners**		**No. utterances**	**Duration**	
	Male	**Female**	**average**	**std**		**average**	**std**
Czech	6	4	28.8	7.4	720	4.45	0.89
English	6	4	26.2	3.6	622	3.57	0.85
French	6	4	26.5	6.8	562	3.72	0.66
German	6	4	26.4	7.1	601	3.51	1.16
Italian	4	6	29.5	6.1	568	3.71	1.13
Polish	3	7	31.8	10.2	631	2.80	1.03
Russian	5	5	23.9	1.7	548	2.85	1.43
Slovakian	3	2	31.6	5.5	573	3.28	1.20
Spanish	5	5	29.3	6.2	551	2.49	1.16
Total	44	41	28.2	6.1	5376	3.38	0.23

Table 7.2: Results of subjective tests for Czech language (average: 41.33%).

	Happiness	**Fear**	**Sadness**	**Anger**	**Neutral**	**Surprise**	**Other**
Happiness	**45**	1	3	1	40	5	5
Fear	0	**29**	11	8	31	11	10
Sadness	0	10	**32**	7	37	8	6
Anger	2	5	5	**55**	16	12	5
Neutral	2	7	8	4	**66**	9	4
Surprise	5	7	8	7	47	**21**	5

Table 7.3: Results of subjective tests for Slovak language (average: 35.33%).

	Happiness	**Fear**	**Sadness**	**Anger**	**Neutral**	**Surprise**	**Other**
Happiness	**20**	0	0	0	71	4	5
Fear	0	**5**	0	0	90	0	8
Sadness	2	4	**36**	5	39	0	9
Anger	0	0	1	**36**	40	16	7
Neutral	0	2	4	5	**84**	5	0
Surprise	2	2	3	2	57	**31**	3

Table 7.4: Results of subjective tests for Polish language (average: 44.66%).

	Happiness	**Fear**	**Sadness**	**Anger**	**Neutral**	**Surprise**	**Other**
Happiness	**59**	2	2	2	24	6	5
Fear	2	**41**	11	13	16	9	8
Sadness	6	14	**36**	7	21	8	9
Anger	5	7	7	**46**	18	6	10
Neutral	12	5	7	5	**51**	9	11
Surprise	13	7	3	12	22	**35**	9

Table 7.5: Results of subjective tests for Russian language (average: 56.5%).

	Happiness	**Fear**	**Sadness**	**Anger**	**Neutral**	**Surprise**	**Other**
Happiness	**70**	6	7	2	6	5	5
Fear	0	**42**	9	18	14	10	7
Sadness	0	12	**50**	13	14	7	4
Anger	0	6	7	**54**	17	13	3
Neutral	4	5	6	11	**64**	7	3
Surprise	2	7	4	11	12	**59**	5

Table 7.6: Results of subjective tests for Spanish language (average: 49.3%).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**	**Other**
Anger	**53**	5	1	21	8	7	5
Fear	9	**14**	2	40	22	10	3
Happiness	0	2	**64**	27	3	3	1
Neutral	2	5	2	**77**	5	5	4
Sadness	9	17	0	25	**40**	8	1
Surprise	5	5	2	32	4	**48**	6

Table 7.7: Results of subjective tests for German language (average: 50.3%).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**	**Other**
Anger	**56**	8	2	19	7	7	1
Fear	11	**30**	4	28	17	8	3
Happiness	2	0	**71**	20	5	1	1
Neutral	5	6	7	**61**	15	5	2
Sadness	2	4	7	34	**43**	7	3
Surprise	10	5	4	26	12	**41**	2

Table 7.8: Results of subjective tests for French language (average: 52%).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**	**Other**
Anger	**69**	8	1	12	4	6	0
Fear	12	**25**	1	27	19	16	2
Happiness	5	1	**69**	14	2	9	0
Neutral	7	6	4	**72**	6	5	0
Sadness	22	22	1	25	**24**	6	0
Surprise	16	7	1	18	5	**53**	1

Table 7.9: Results of subjective tests for Italian language (average: 46.3%).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**	**Other**
Anger	**69**	4	1	13	5	5	2
Fear	0	**0**	0	0	0	0	0
Happiness	5	1	**71**	12	2	3	7
Neutral	6	4	**4**	70	9	5	3
Sadness	15	12	0	16	**46**	5	5
Surprise	7	0	0	52	11	**22**	8

Table 7.10: Results of subjective tests for English language (average: 50.4%).

	Anger	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**	**Other**
Anger	**63**	5	0	25	3	4	0
Fear	3	**25**	0	58	0	14	0
Happiness	5	5	**55**	20	0	15	0
Neutral	9	7	**7**	68	3	6	2
Sadness	13	17	1	59	**4**	1	4
Surprise	14	7	1	37	0	**38**	3

7.2 Regression based emotion recognition from Slavic speech

This experiment was carried out in time when only speech of Slavic languages was available. However, the results can be generalized for all available databases. The aim of this experiment is to propose an algorithm for emotion recognition with the capability of mapping the output emotion into two-dimensional space of emotions. The results reported in this section were published in [ASE12].

7.2.1 Feature extraction

Feature extraction process is done as follows:

1- Speech signal is segmented into frames of 32ms with 50% overlap.

2- Features in Table 7.11 are extracted from each frame.

3- High-level characteristics are computed from segmental features obtained from the previous step (Table 7.12).

4- High-level characteristics are concatenated into the final feature vector used for training.

Beside the features listed in Table 7.11, the first and second differences (Δ, $\Delta\Delta$) of these features were also considered.

Table 7.11: List of features for emotion recognition from Slavic speech.

Feature	**Abbrev.**	**Number of coefficients**
Mel Frequency Cepstral Coefficients	MFCC	20
Human Factor Cepstral Coefficients	HFCC	20
Linear Frequency Cepstral Coefficients	LFCC	20
MEL Bank Spectral Coefficients	MELBS	20
Human Factor Bank Spectral Coefficients	HFBS	20
Linear Frequency Bank Spectral. Coefficients	LFBS	20
Perceptual Linear Predictive 1	PLP1	21
Perceptual Linear Predictive 2	PLP2	21
Perceptual Linear Predictive 3	PLP3	21
Perceptual Linear Predictive 4	PLP4	11
4Hz Modulation Energy	4HzME	1
Mel Spectrum Modulation Energy	MSME	1
Fundamental frequency	F0	1
Formant frequencies	Fx	5
Formant Bandwidths	Bx	5
Harmonicity	H	1
Temporal Energy	TE	1

Teager Energy Operator	TEO	1
Zero Crossing Ratio	ZCR	1

The total number of features ξ_{all} extracted from each utterance is computed as follows

$$\xi_{\text{all}} = \xi_{\text{coef}}\xi_{\text{high}}\xi_{\text{wave}} = 211\ .34\ .3 = 21522 \tag{7.1}$$

Where ξ_{coef} is the total number of feature coefficients, ξ_{high} is the number of high-level characteristics and ξ_{wave} is the number of feature waveforms (one original and two differences)

Table 7.12: List of suprasegmental (high-level) features extracted from segmental features for emotion recognition from Slavic speech.

Basic characteristics	mean, median, standard deviation, maximum, minimum, range, slope
Positional characteristics	position of maximum, position of minimum
Relative characteristics	relative standard deviation, relative range, relative maximum, relative minimum, relative position of maximum, relative position of minimum
moments	kurtosis, skewness, Pearson's skewness coefficient, 5th moment, 6th moment
Regression characteristics	linear regression coefficient, linear regression error
percentiles	1%, 5%, 10%, 20%, 30%, 40%, 60%, 70%, 80%, 90%, 95% and 99% percentile

7.2.2 Feature selection

The feature selection process depends on the regression algorithm used. However, for all regression techniques the minimum Redundancy Maximum Relevance (mRMR) algorithm is employed in order to reduce the number of features from 21522 to 200. As it will be shown in the next section, one of the regression techniques is combined with a forward feature selection.

7.2.3 Regression

Several regression algorithms were tested in order to identify the best one among them. The considered algorithms are the following

1. Feedforward neural network with one input, two hidden and one output layers (ANN).
2. Support vector regression with linear kernel (SVR-LK)
3. Support vector regression with radial basis kernel (SVR-RBF)
4. Support vector regression with radial basis kernel combined with feature forward selection (SVR-RBF-FS).

The speech corpus was split into three parts, where 80% of this corpus was used for training, 10% for validation and 10% for testing. The mean absolute errors for all regression techniques are reported in Table 7.13, where the maximum possible error for each emotion is 10, which corresponds to the number of listeners who were involved in the subjective evaluation process for Czech, Polish and Russian. The subjective evaluation results for Slovakian language were doubled in order to have all emotions in the same scale.

Table 7.13: Mean absolute errors for regression algorithms under examination.

	anger	happiness	neutral	sad	surprise
ANN	4.35	3.84	3.53	2.86	3.14
SVR-LK	2.01	1.78	2.42	1.77	1.94
SVR-RBF	1.93	1.69	2.30	1.69	1.81
SVR-RBF-FS	1.10	0.87	1.79	0.86	1.05

The results suggests that the support vector regression with radial basis kernel combined with feature forward selection (SVR-RBF-FS) gives the best result among the other algorithms. Hence, the next is devoted to give more details about SVR-RBF-FS method. The MAEs for all emotions using this method are shown in Figure 7.1.

The regression is carried out as follows: First, as it was stated in the feature selection section, the mRMR algorithms is applied resulting in 200 features, these features are subsequently filtered by using forward selection, the selection criteria is the Mean Absolute Error (MAE) defined as

$$E_{\mathrm{mae}} = \frac{1}{MN}\sum_{c=1}^{M}\sum_{i=0}^{N-1} |(s_c^i - y_c\ )| \quad , \tag{7.2}$$

where:

- M is the number of classes (emotions), M=5.
- N is the number of patterns (speech stimuli) used for validation, N=280.
- y is the output of SVR, y $= \{y_1, y_2, \dots, y_M\ \}$. $y \in \mathbb{R}^M$.
- s is the vector of subjective evaluation results for the i^{th} stimulus, $s \in \mathbb{N}^M$.

The forward feature selection is described in the following pseudocode

```
SET $Z_0 = \emptyset$          //output feature group
SET $m = 1$                    //feature index
SET $J(0) = 0$                 //mean absolute  error function initialization
SET $N_{\mathrm{fs}} = 20$     //number of iterations
SET $N_{\mathrm{f}} = 200$     //number of features considered
SET $N \ = 280$                //number of feature vectors
SET $M \ = 5$                  //number of classes (emotions)

//the main cycle of  FS algorithm
FOR $m$=0 TO $N_{\mathrm{fs}} - 1$

    FOR $n$=1 TO $N_{\mathrm{f}}$          // adding features

        FOR  $i$= TO 10        // 10-fold cross validation
        //regression of the $i^{th}$ feature  vector  $\mathbf{F}_i$ by using SVR-RBF

                $y_c = C(\mathbf{F}_i(Z_{\boldsymbol{m}} \cup \ n))$

                $E_{\mathrm{mae}}^{n} = \frac{1}{M}\sum_{c=1}^{M} |(s^i \ - y \ )|$    //mean absolute  error

        ENDFOR  // end of validation cycle

    ENDFOR // end of feature adding cycle

 $f^{+} = \mathrm{argmin} \ \ E_{\mathrm{mae}}^{n}$      // find feature with minimal MAE
 $Z_{\boldsymbol{m+1}} = Z \cup f^{+}$          // add the selected feature to group Z
 $J(m+1) = E_{\mathrm{mae}}^{f+}$       // update J

ENDFOR   // end the main cycle of  FS algorithm
```

The forward selection resulted in 43 features for which the MAE was minimal, these features are reported in Table 7.14, where the first column contains the coefficient number from which the high-level feature was extracted, the second column represents the feature group and the third column contains the high-level feature. The last column indicates whether the high-level feature was extracted from the original feature pattern (O) or from its first (Δ) or second derivative ($\Delta\Delta$).

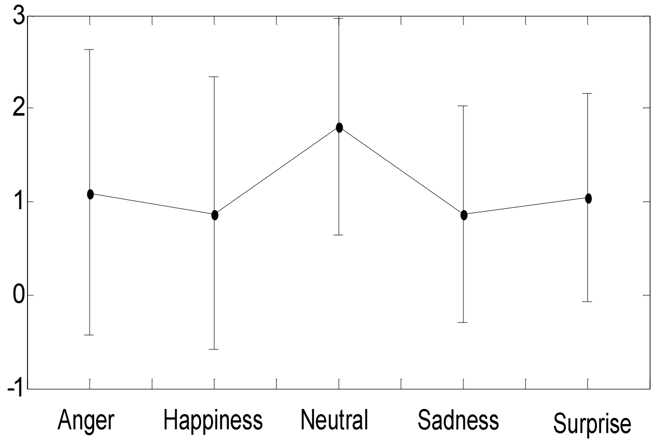


Figure 7.1: Mean Absolute errors for SVR-RBF-FS algorithm.

Table 7.14: List of selected features using SVR-RBF-FS.

Coef.	Feature	High-level feature	P	Coef.	Feature	High-level feature	P
8	MFCC	Slope	O	9	MFCC	70% percentile	Δ
13	PLP3	Relative minimum	Δ	9	MELBS	20% percentile	O
11	PLP4	10% percentile	Δ	11	PLP4	5% percentile	O
9	MELBS	30% percentile	O	10	PLP1	Slope	O
12	HFCC	regression coefficient	O	19	LFCC	Slope	O
8	PLP3	minimum	ΔΔ	9	PLP1	10% percentile	ΔΔ
1	HFCC	70% percentile	Δ	8	PLP3	maximum	Δ
9	MFCC	minimum	O	16	PLP1	60% percentile	Δ
1	PLP2	maximum	Δ	20	LFCC	Regression coefficient	ΔΔ
1	HFCC	Range	ΔΔ	16	LFCC	80% percentile	Δ
16	PLP1	80% percentile	Δ	3	MELBS	Mean	O
19	LFCC	maximum	ΔΔ	19	HFCC	minimum	Δ
18	LFCC	slope	Δ	4	PLP3	maximum	Δ
1	HFCC	80% percentile	Δ	9	PLP1	slope	O
16	PLP3	70% percentile	O	15	PLP3	20% percentile	ΔΔ
8	MFCC	80% percentile	O	15	PLP2	regression coefficient	O
16	PLP1	mean	Δ	8	MFCC	relative standard dev.	Δ
18	LFCC	minimum	O	1	HFCC	range	O
1	HFCC	percentile1	Δ	16	PLP1	90% percentile	Δ
1	PLP1	regression coefficient	Δ	12	PLP1	slope	Δ
19	LFCC	range	ΔΔ	14	PLP1	regression coefficient	O

7.2.4 Mapping emotions onto two-dimensional space

The Majority of research work addressing emotion recognition from speech has focused on the classical approach to the task; the input speech signal is assigned to one class "emotion' according to a certain classification criteria, for example, the logarithmic likelihood. This approach has one shortage; because the output of such systems is determined within discrete emotions whereas it is proved that human emotional states are characterized by a high level of variability [Pic00]. Moreover, it is impossible to build a speech database that covers all human emotional states.

Some studies suggested approaches that estimate the activation and valence values of speech emotions. Such systems are basically trained using emotional corpora labeled for this purpose, where the listeners are asked to guess the position of each stimulus under examination in the two dimensional emotional space defined by the activation and valence axes (Figure7.2) [Osg57]. However, this kind of labeling is more time consuming and requires experts or people with a special training.

In fact, most speech emotional databases proposed so far consider only a certain number of emotions. For example, BDES database contains speech stimuli of seven emotional states: Anger, boredom, fear, happiness, disgust and neutral.

In the light of remarked above, it is obvious that an approach where discrete classes (emotions) are used to estimate the activation and valence values can be very useful, because it combines the simplicity of constructing emotional databases of discrete emotions with the advantages of the continues representation of emotions.

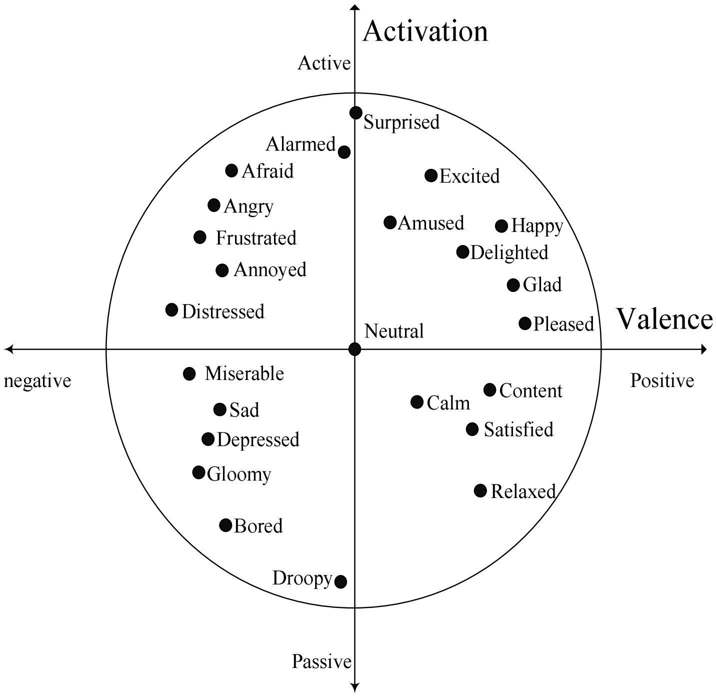


Figure 7.2: two-dimensional space of emotions.

In the next a new approach called skeleton is presented for mapping discrete emotion into two-dimensional space of emotions. Taking into account the current experiment with emotion recognition from Slavic speech, the skeleton approach works according to the procedure describe below.

After identifying the optimal features, the regression is provided by using regression function C as following: the feature vector $\mathbf{F}$ is extracted from the input speech signal. After that, only optimal features Z_{opt} selected using the forward selection algorithm are considered.

$$\mathrm{y} = C\left(\mathbf{F}(Z_{\mathrm{opt}})\right) \tag{7.1}$$

Ratios d_{ij} defined as (7.2) are computed for all possible pairs of emotional states, these ratios determine the positions of points b_{ij} in the two dimensional emotional space. These points are always located on the connecting lines between what is called "fixed points". For example, b_{13} is located between anger and neutral and b_{25} is located between happiness and sadness.

$$d_{ij} = \frac{y_i}{y_i + y_j}, \tag{7.2}$$

in case of $y_i = y_j = 0$, $d_{ij}$is not considered.

The fixed points represent the positions of emotions considered in the experiment; these positions were set according to the activation-valence theory proposed in [MR74]. The fixed point coordinates are shown in Table 7.15.

Table 7.15: The coordinators of fixed points.

	Valence (V)	**Activation (A)**
Anger	-1	1
Happiness	1	1
Neutral	0	0
Sadness	-1	-1
Surprise	0	1

For emotions i,j the coordinates of b_{ij} are computed as

$$b_{ij} = \left(1 - d_{ij}\right)\left(V_j, A_j\right) + d_{ij}(V_i, A_i), \tag{7.3}$$

where V_j, A_j, V_i, A_i are the valence and activation values of fixed points j and i respectively.

Finally, the centorid of b_{ij} points is computed, this centroid determines the position of the input speech emotion in the two-dimensional space.

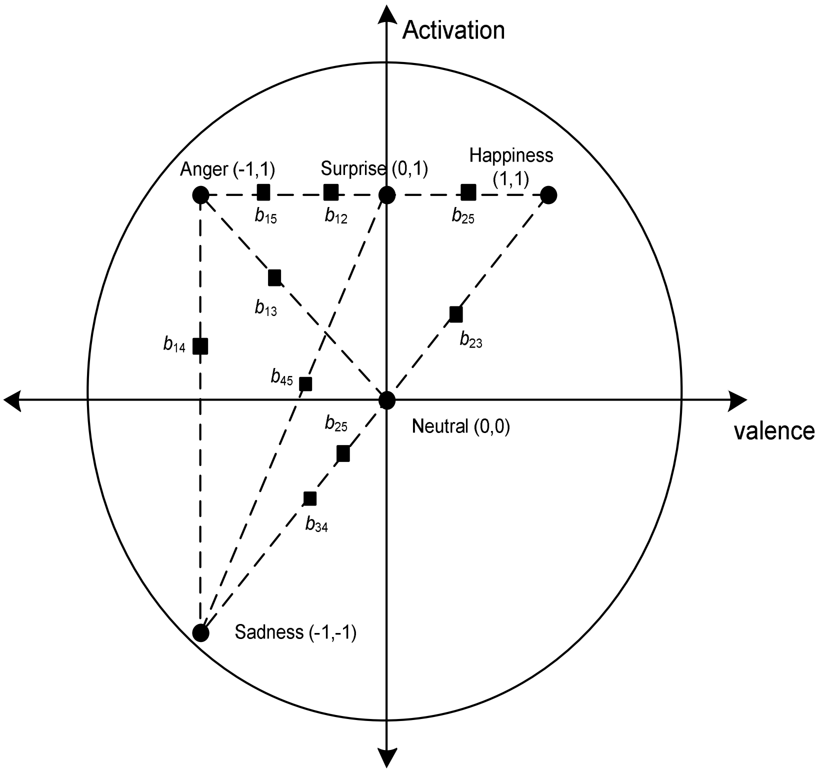


Figure 7.3: Two-dimensional emotional space with fixed and *b* points.

As an example, suppose that the regression algorithm returned vector $y = \{6,0,2,2,0\ \}$. As it was mentioned before, this vector contains the outputs for each emotion in the following order: anger, happiness, neutral, sadness and surprise. The ratio d_{13} between anger and neutral is computed as

$$d_{13} = \frac{y_1}{y_1 + y_3} = \frac{6}{6+2} = 0.75 \tag{7.4}$$

The remaining ratios are computed analogously:

d_{12}=d_{15}=d_{35}=d_{45}=1, d_{13}=d_{14}=0.75, d_{23}=d_{24}=0, d_{34}=0.5, d_{25} is not considered.

Now the aim is to find the position of b_{13} according to (7.3)

$$b_{13} = (1 - 0.75)(0,0) - 0.75(-1,1) = (-0.75,0.75) \tag{7.5}$$

Again, the remaining points are taken analogously:

b_{12}=(-1,1), b_{13}=(-0.75,0.75), b_{14}=(-1,0.5), b_{15}=(-1,1), b_{23}=(0,0) , b_{24}=(-1,-1), b_{34}=(-0.5,-0.5), b_{35}=(0,0), b_{45}=(-1,-1)

The emotion position is then determined by getting the centroid of points b_{ij}, which is (-0.694, 0.083), this means that the input speech has high negative valence and low positive activation.

The secondary evaluation was carried out by 5 listeners, who labeled a small corpus of 50 emotional speech utterances according to activation-valence protocol. These listeners were asked to guess the position of speakers' emotional states from 50 utterances. The same utterances were analyzed by the proposed algorithm and the results of subjective evaluation and automatic recognition were compared in terms of MAE. The mean absolute errors for valence and activation obtained by the two-dimensional approach are reported in Table 7.16. The maximum possible error here is 2, which the difference between the minimum (-1) and maximum (+1) values of valence and activation.

Table 7.16: Mean absolute errors for both valence and activation using skeleton approach.

Valence	0.37
Activation	0.39

As a practical example, Figure 7.4 illustrates the results of emotional analysis over client-operator phone call; the first column illustrates the two-dimensional emotional space of each channel whereas the second column represents the activation and valence values in time domain. The big advantage of the two-dimensional representation is that it gives a very good overview of the speaker's emotions distribution within the utterance as well as the intensities of these emotions. However, the basic form of such interpretation doesn't contain any temporal information, that is, it is not possible to define when a concrete emotion occurs. This limitation can be avoided by displaying the activation and valence in time (like the right column of Figure 7.4) or by using a three-dimensional interpretation, where the third dimension is time.

7.2.5 Summary

In this section, a new approach for automatic recognition of emotional expressions from spontaneous Slavic speech was proposed; this approach is based on Support vector regression with radial basis function combined with forward selection of features. Moreover, a new method for the mapping of discrete emotions into continuous two-dimensional space was presented. The results of experiments made are promising; the SVR-RBF-FS yielded remarkably good performance for all emotions, where the MAE was 1.134±0.381.

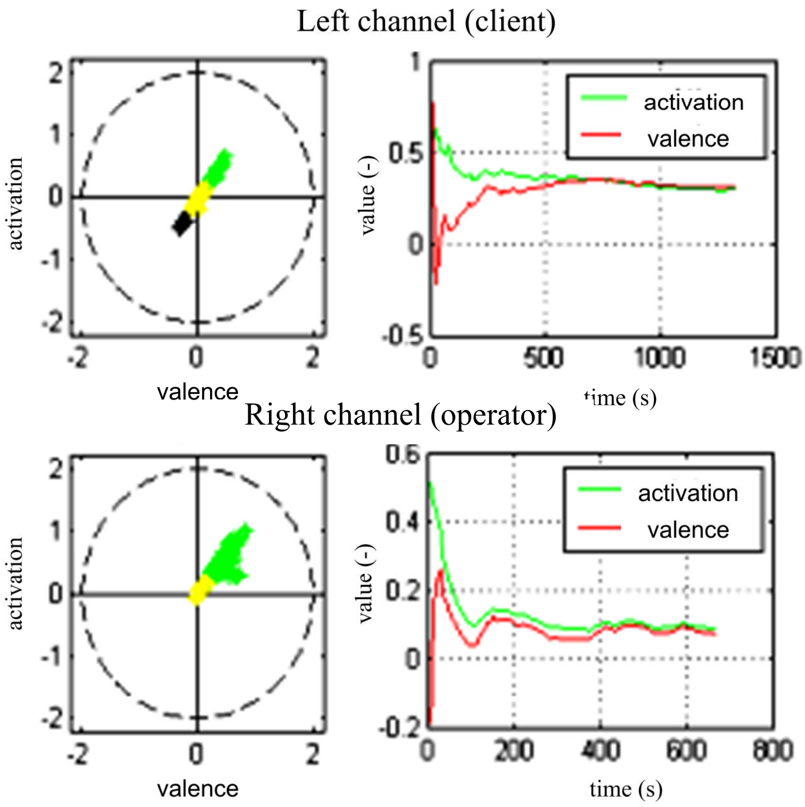


Figure 7.4: Results of emotional analysis over client-operator phone call.

7.3 Multilingual system for emotion recognition from spontaneous speech obtained from call centers

The aim of this section is to propose a global system for emotion recognition using all languages available in MSDES.

7.3.1 Adaptation of Speech database

As it was mentioned above, Multilingual Spontaneous Database of Emotional Speech is employed in this section. However, only utterances which achieved 80% score in subjective evaluation tests were considered. This is due to the fact that comparing to the previous experiment which employed a regression technique the approach presented in this section will be a classification task. The following notices should be mentioned before the system description.

- The fear state was discarded due to the lack of sufficient speech with this emotion and moreover, since the emotion recognition system will be used in call centers which focus on telemarketing and customer care services, there was no interest in this emotion.

- It was found that the surprise state can't be considered as an independent state, because speakers can express surprise in both positive and negative way. Thus it will be processed independently.
- The anger and happiness classes were split into two classes; one with low activation level and one with high activation level. The anger with low activation level can approximately represent annoy emotional state where as the low level happiness represents the please emotion (see Figure 7.3).

The numbers of utterances used in further experiments for each state are reported in Table 7.17.

Table 7.17: Number of utterances used for emotion recognition using MSDES.

Neutral		Anger L		Anger H		Happ. L		Happ. H		Sadness		Surprise	
M	F	M	F	M	F	M	F	M	F	M	F	M	F
286	269	127	93	150	158	118	126	55	92	102	155	107	102

7.3.2 Preprocessing and feature extraction

The telephone records in call centers usually have two channels, where the first channel contains the speech of the agent whereas the second channel contains client's speech. These channels are processed separately. Each channel is segmented into supersegments (Figure 3.7) with length of 4.096 seconds (32768 samples for sampling frequency 8kHz) with 50% overlap. The length of the supersegment satisfies the condition of using FFT in further feature extraction and is close to the average length of training samples in MSDES reported in Table 7.1. Moreover, this length is neither too short to not capture the speaker's emotional state nor too long to reduce the systems temporal resolution. Each supersegment is subsequently segmented into 64ms long segments with 50% overlap.

The biggest set of features was put under examination in section 6.3, where 25 different kinds of acoustic features were examined (Table 6.8). However ACW and LPCC features didn't show any significant discrimination power within emotions for both acted and spontaneous databases and thus they will not be considered in the following experiments. Moreover, despite that Wavelet Decomposition WADE showed a good performance in terms of distinguishing between emotional states. It was decided to discard them as well due to their computational complexity. The features considered for emotion recognition using MSDES are reported in Table 7.18.

Table 7.18: List of features used for emotion recognition using MSDES.

Feature	Abbreviation	Number of coefficents
Mel Freq. Cepstral Coefficients	MFCC	20
Human Factor Cepstral Coefficients	HFCC	20
Linear Frequency Cepstral Coefficients	LFCC	20
Perceptual Linear Predictive 1	PLP1	21
Perceptual Linear Predictive 2	PLP2	21
Perceptual Linear Predictive 3	PLP3	21
Perceptual Linear Predictive 4	PLP4	11
MEL Bank Spectral Coefficients	MELBS	20
Human Factor Bank Spectral Coefficients	HFBS	20
Linear Frequency Bank Spectral Coefficients	LFBS	20
Linear Predictive Coefficients	LPC	10
Mel Spectral Modulation Energy	MSME	1
4Hz Modulation Energy	4HzME	1
Spectral Features	SF	5
Fundamental frequency	F0	1
Formant frequencies	Fx	5
Formant Bandwidths	Bx	5
Harmonicity	H	1
Temporal Energy	TE	1
Teager Energy Operator	TEO	1
Zero Crossing Ratio	ZCR	1
Total		225

7.3.3 General classifier

First, the classification task is approached in a classical manner, which means that several classifiers are employed with segmental, suprasegmental features and the fusion of both types of these feature representations.

Two methods for fusing segmental and suprasegmental features are proposed (Figure 7.5)

1. **Fusion on feature level**: The fusion is applied aiming to create a single feature vector from both segmental and suprasegmental features. In this case, both vectors were concatenated in order to get the final feature vector.
2. **Fusion on classifier level**: Two different classifiers are used to classify segmental and suprasegmental feature vectors. This approach was previously presented in section 6.2. The output scores of each classifier are multiplied analogously to Figure 6.5. More on combining pattern classifiers can be found in [Kun04].

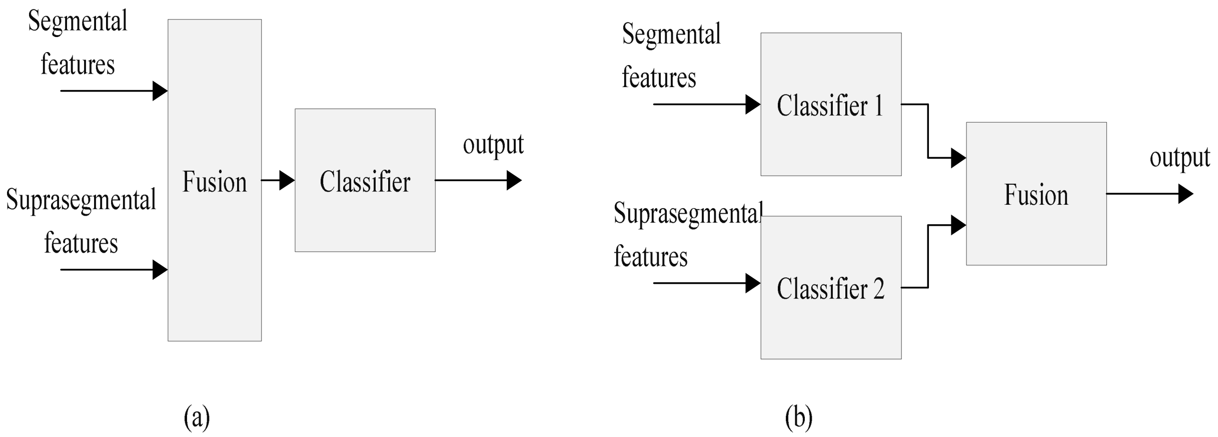


Figure 7.5: Two possible ways of fusing segmental and suprasegmental features. Fusion on feature level (a) and fusion on classifier level (b).

Table 7.19 contains the results of different classifiers and feature selection methods for suprasegmental features whereas Table 7.20 reports the results for segmental features. The classification accuracies obtained by fusing both types of features are shown in Table 7.21.

Table 7.19: Weighted Classification accuracies using suprasegmental features.

	SVM-RBF	SVM-linear	GMM	DT
mRMR (50)	60.26	55.54	55.65	51.65
mRMR+SFFS	62.45	59.26	56.28	54.48
mRMR+FS	61.25	55.54	55.65	53.34
mRMR+BS	61.57	55.75	56.35	54.68

Table 7.20: Weighted Classification accuracies using segmental features.

	SVM-RBF	SVM-linear	GMM	DT
TMV	61.04	57.41	56.78	52.62
PCA	49.41	45.30	46.13	44.60
k-means	51.74	49.31	54.35	50.05

Table 7.21: Classification results for different types of fusion of segmental and suprasegmental features.

Fusion level	Segmental	Suprasegmental	Result (%)
Feature level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	63.178
Classifier level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	59.632

Results in tables 7.19 and 7.20 suggests that support vector machines classifier with features selected by combining mRMR and SFFS algorithm gives best classification accuracy (62.45%) for suprasegmental features. Regarding segmental features, the best result is achieved by applying Temporal Mean Vector (TMV) for feature reduction and again, SVM-RBF classifier shows the best performance.

The fusion results reported in Table 7.21 indicate that there is a slight improvement when the feature vectors are fused on feature level. On the other hand, the fusion on classifier level yields decreasing the classification accuracy comparing to the case when only suprasegmental features are used.

7.3.4 Gender-dependent system

According to relevant literature and own experiments it is proven that gender-dependent approach to the recognition of vocal emotions can deliver better classification accuracy comparing to gender-independent approach. Hence, this section reports the results of emotion recognition for both male and female speakers.

Table 7.22: Weighted Classification accuracies for male speakers using suprasegmental features.

	SVM-RBF	**SVM-linear**	**GMM**	**DT**
mRMR (50)	63.24	59.54	59.11	54.63
mRMR+SFFS	66.71	61.88	60.28	58.42
mRMR+FS	66.71	61.01	60.28	60.39
mRMR+BS	65.63	61.55	61.635	61.28

Table 7.23: Weighted Classification accuracies for male speakers using segmental features.

	SVM-RBF	**SVM-linear**	**GMM**	**DT**
TMV	62.0449	61.54	61.78	58.62
PCA	49.41	50.68	52.13	47.60
k-means	51.74	49.31	54.35	50.05

Table 7.24: Classification results for different types of fusion of segmental and suprasegmental features for male speakers.

Fusion level	**Segmental**	**Suprasegmental**	**Result (%)**
Feature level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	67.42
Classifier level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	67.86

Table 7.25: Classification accuracies for female speakers using suprasegmental features.

	SVM-RBF	SVM-linear	GMM	DT
mRMR (50)	57.14	55.84	56.20	56.91
mRMR+SFFS	59.65	59.12	58.36	56.49
mRMR+FS	58.17	59.30	58.07	54.33
mRMR+BS	59.31	59.55	58.18	58.40

Table 7.26: Classification accuracies for female speakers using segmental features.

	SVM-RBF	SVM-linear	GMM	DT
TMV	55.14	54.57	53.48	50.98
PCA	41.46	38.42	38.83	38.30
k-means	51.5	47.3	47.71	44.63

Table 7.27: Classification results for different types of fusion of segmental and suprasegmental features for female speakers.

Fusion level	Segmental	Suprasegmental	Result (%)
Feature level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	59.90
Classifier level	TMV+SVM-RBF	mRMR+SFFS+SVM-RBF	59.13

For gender-dependent approach, the results reported in Tables 7.22 to 7.27 proves that the combination of mRMR and SFFS algorithms for feature selection and SVM-RBF classifier works best for both male and female speakers. The fusion of segmental and suprasegmental features didn't reveal any significant improvement comparing to the case when only suprasegmental features are used. Thus it was decided to discard segmental features in further experiments.

7.3.5 Emotion coupling

Despite that this approach was validated on acted speech only with GMM classifiers, an arbitrary classifier can be used as a classification element in this approach. The first task then is to find the best classification element for emotion coupling that best fits MSDES. Table 7.28 reports results of four different classifiers using suprasegmental features and mRMR for feature selection. The results suggest that SVM-RBF classifier works best for emotion coupling. The confusion matrices for SVM-RBF with mRMR and SVM-RBF with mRMR+SFFS are reported in Tables 7.29 and 7.30 respectively.

Table 7.28: Classification accuracies for different classification elements of emotion coupling approach.

SVM-RBF	SVM-Linear	GMM	DT
61.8096	59. 6999	56.9509	55.6570

Table 7.29: Confusion matrix for emotion coupling using SVM-RBF with mRMR (average classification accuracy: 61.8096%)

	Anger L	Anger H	Happiness H	Happiness L	Neutral	Sadness
Anger L	**57.7778**	24.4444	6.1111	1.1111	7.7778	2.7778
Anger H	10.4530	**81.5331**	4.5296	0	3.4843	0
Happiness H	8.3929	7.5893	**47.5893**	0.4464	27.9464	8.0357
Happiness L	13.4211	5.7895	13.6842	**45.5263**	20.2632	1.3158
Neutral	5.2252	1.4414	4.5045	0.1802	**82.1622**	6.4865
Sadness	4.9751	0.9950	5.4478	0	32.3134	**56.2687**

Table 7.30: Confusion matrix for emotion coupling using SVM-RBF with mRMR (average classification accuracy: 67.4135%).

	Anger L	Anger H	Happiness H	Happiness L	Neutral	Sadness
Anger L	**64.4444**	21.1111	3.3333	0	8.8889	2.2222
Anger H	10.4530	**83.6237**	2.0906	0	3.4843	0.3484
Happiness H	10.6250	5.3571	**56.0714**	0	18.1250	9.8214
Happiness L	7.1053	7.1053	11.5789	**56.3158**	17.8947	0
Neutral	5.7658	1.0811	6.6667	0	**79.2793**	7.2072
Sadness	7.9602	0.4975	5.9453	0	20.8507	**64.7463**

7.3.6 Fusion of all systems

Outputs from all classifiers namely general, emotion coupling and gender-dependent are fused using one layer perceptron with 50 neurons. The outputs of the fusion layer are connected to the two-dimensional mapping layer except the surprise which is, as it was stated before, processed independently. The final confusion matrix obtained by fusing all systems is reported in Table 7.31 and the comparison of results for several feature extraction, reduction and classification techniques for MSDES are summarized in Figure 7.6. The block scheme of the complex classification system is illustrated in figure 7.7.

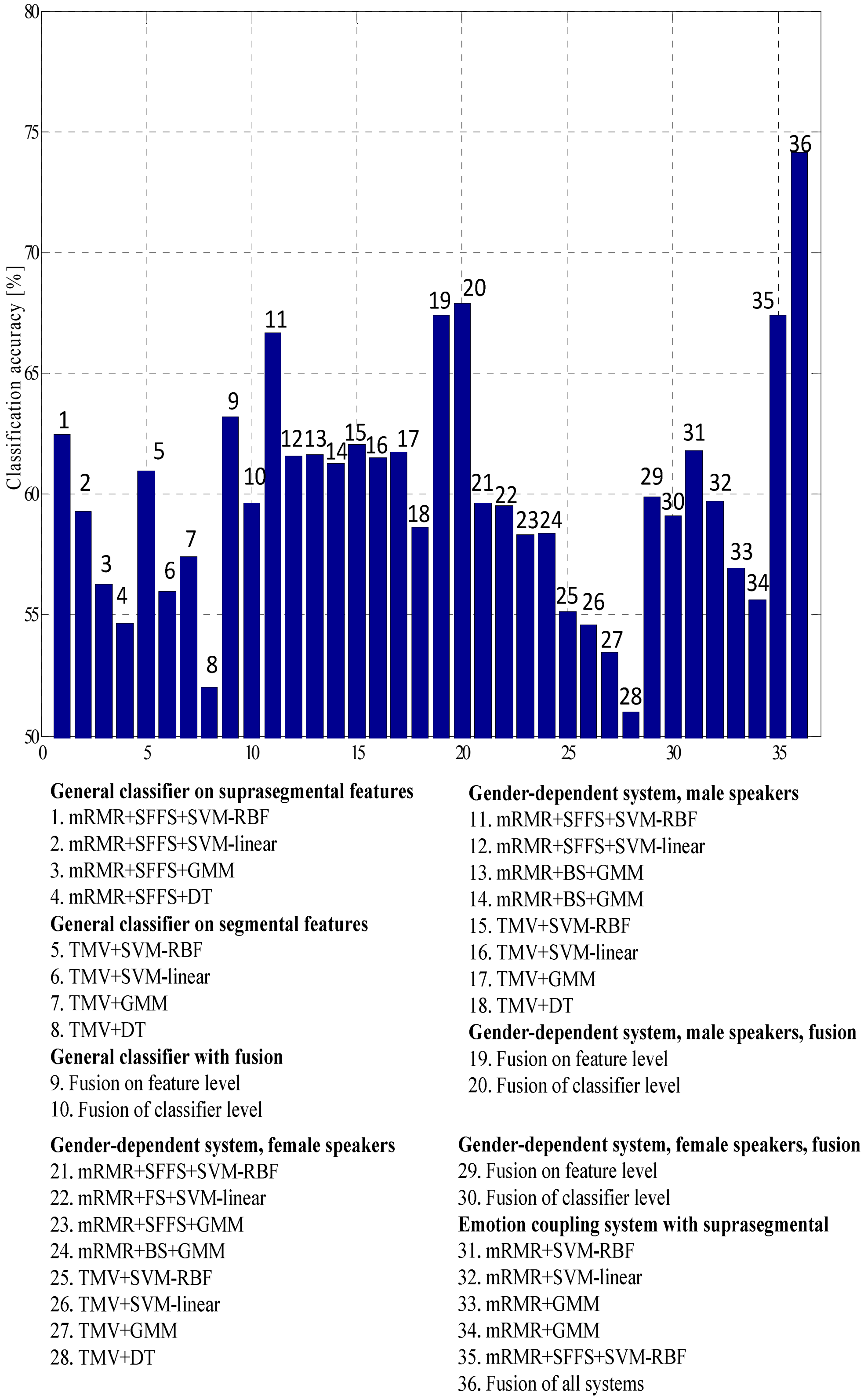


Figure 7.6: Comparison of results for several feature extraction, reduction and classification techniques for MSDES.

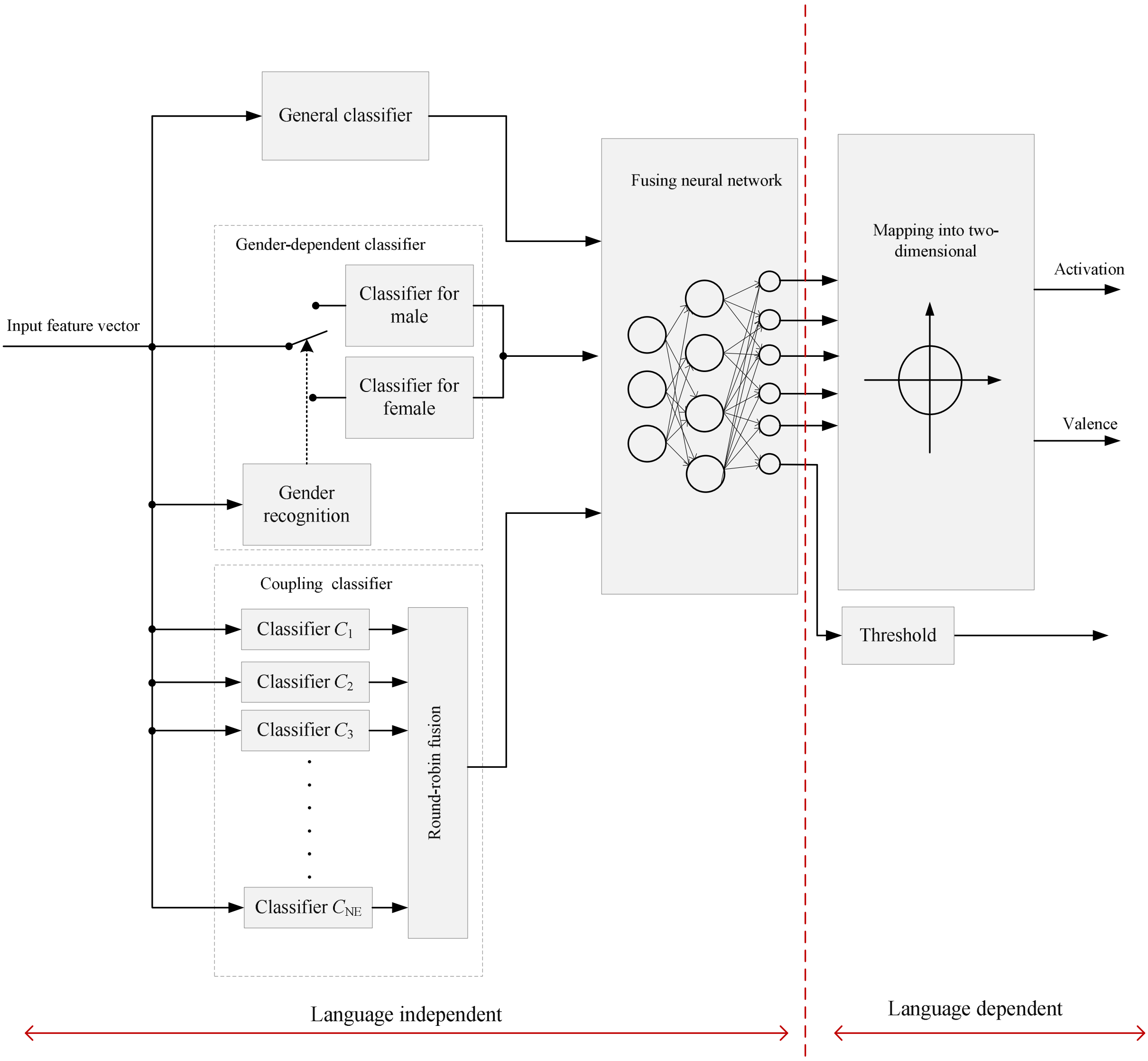


Figure 7.7: Block scheme of proposed system for emotion recognition form MSDES.

Table 7.31: Final confusion matrix obtained after fusing all classifiers (average classification accuracy 74.16%).

	Anger L	Anger H	Happiness L	Happiness H	Neutral	Sadness
Anger L	76	8	2	2	5	7
Anger H	6	76	2	14	0	2
Happiness L	7	3	73	10	7	0
Happiness H	2	13	4	70	4	2
Neutral	4	0	4	4	78	10
Sadness	9	3	4	2	10	72

7.3.7 Mapping emotions into two-dimensional space

The last step within the proposed system is the mapping of fused outputs of all employed classifiers into the two-dimensional emotional space.

1. Skeleton method: This previously introduced method was tested with the following fixed points

Table 7.32: Proposed fixed points of skeleton approach for MSDES.

	Valence (V)	Activation (A)
Anger L	-1	1
Anger H	-0.5	0.5
Happiness L	1	1
Happiness H	0.5	0.5
Neutral	0	0
Sadness	-1	-1

2. Neural network approach: This approach requires training patterns subjectively evaluated according to the active-valence protocol. The same set of training data used for the evaluation of method presented in section was employed. However, it should be stated that this small set includes only Czech utterances.

The results of two-dimensional mapping in terms of mean absolute errors for both skeleton and neural network approaches are reported in Table 7.33.

Table 7.33: Evaluation of different approach for two-dimensional mapping in terms of MAE.

	Skeleton approach	2D trained ANN
MAE for valance	0.42	0.21
MAE for activation	0.40	0.26

7.3.8 Detection of surprise state

The detection of surprise state is carried out by using SVM classifier with linear function, as it showed the best classification performance among other classifiers. For each supersegment, this classifier is applied, if the classifier output exceeds a certain threshold, then surprise is registered beside the original emotion. Figure 7.8 illustrates ROC curves of several classifiers for the detection of surprise state whereas the detailed results of these classifiers are reported in Table 7.34.

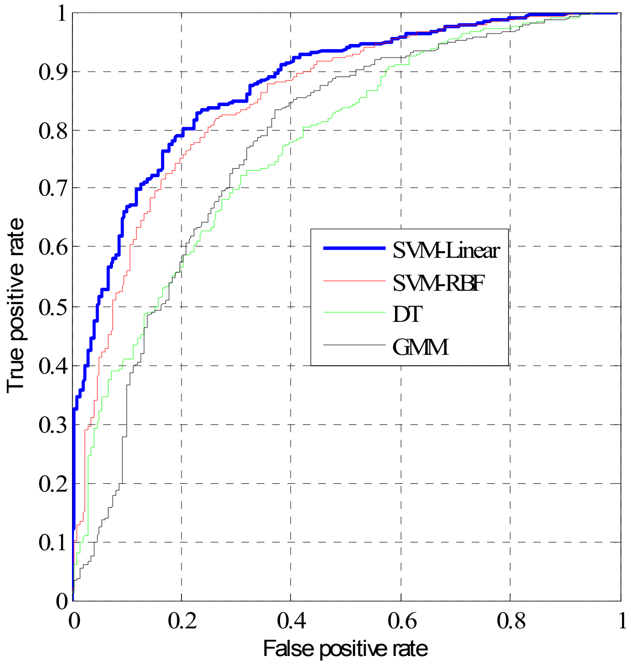


Figure 7.8: ROC curves of different classifiers for the detection of surprise state.

Table 7.34: Results of surprise detection for several classifiers.

	SVM-RBF	**SVM-linear**	**NBC**	**DT**
Weighted accuracy	81.11%	83.51%	77.26%	81.25%
Unweighted accuracy	70.67%	75.07%	75.09%	73.54%
Precision	92.61%	92.79%	79.63%	89.72%
Recall	83.57%	85.97%	88.40%	85.56%
F-measure	87.86%	89.25%	83.79%	87.59%
Matthews correlation	0.47	0.54	0.46	0.49%

7.3.9 Selected suprasegmental features

The section reports the suprasegmental features selected for all classifiers involved in the complex system.

Table 7.35: List of suprasegmental features selected for gender-dependent system.

Male speakers				Female speakers			
Coef.	**Feature**	**High-level feature**	**P**	**Coef.**	**Feature**	**High-level feature**	**P**
1	PLP3	90% percentile	Δ	14	MFCC	Median	Δ
9	PLP3	90% percentile	O	1	MFCC	20% percentile	O
15	MFCC	30% percentile	Δ	19	LFCC	Slope	O
12	PLP2	Median	Δ	14	PLP2	60% percentile	Δ
17	LFCC	99% percentile	ΔΔ	9	PLP3	90% percentile	O
9	PLP2	Slope	O	16	PLP2	skewness	Δ
18	MFCC	Slope	O	16	MFCC	Percentile range	Δ
3	MFCC	Minimum	Δ	5	PLP1	Slope	Δ
2	HFCC	10 percentile	O	5	PLP2	Relative standard dev	Δ
13	PLP2	30% percentile	O				

Table 7.36: List of suprasegmental features selected for general classifier and classifier for surprise detection.

General system				System for surprise detection			
Coef.	**Feature**	**High-level feature**	**P**	**Coef.**	**Feature**	**High-level feature**	**P**
3	LFCC	30% percentile	O	11	MFCC	Maximum	Δ
13	MFCC	Range	O	10	LFCC	Pearson skewness coeficient	Δ
3	PLP4	70% percentile	O	14	PLP1	Percentile range	O
-	CPP	80% percentile	O	2	HFCC	10% percentile	O
19	MELCS	30% percentile	O	6	MFCC	Regression coefficient	O
-	F0	slope	ΔΔ	17	LFCC	95% percentile	ΔΔ
2	LFBS	5th moment	O	20	LFCC	6th moment	O
5	PLP3	Kurtosis	O	10	HFCC	95% percentile	O
5	PLP4	Slope	Δ	2	HFCC	95% percentile	O
19	LFCC	5% percentile	Δ				
7	MELBS	slope	O				
2	HFCC	95% percentile	O				
8	HFCC	70% percentile	O				
-	SF kurtosis	Pearson skewness coefficient	O				
-	F0	20% percentile	Δ				

Table 7.37: List of suprasegmental features selected for emotion coupling approach.

Coef	**Feature**	**High-level feature**	**P**	**Coef.**	**Feature**	**High-level feature**	**P**
Anger L X Anger H				**Anger H X Happiness L**			
-	F0	10% percentile	O	4	MELBS	Maximum	O
13	MFCC	40% percentile	O	-	CPP	95% percentile	O
5	MFCC	5% percentile	O	17	LFBS	20% percentile	O
-	F0	Slope	O	17	LFCC	Minimum	O
15	LFBS	30% percentile	O	13	HFCC	Range	O
3	LFCC	70% percentile	O	-	F0	10% percentile	Δ
Anger L X Happiness L				8	PLP4	Slope	O
8	PLP2	slope	O	2	PLP2	Skewness	O
9	HFCC	Maximum position	O	2	PLP2	Kurtosis	O
5	LFCC	Maximum position	Δ	**Anger H X Happiness L**			
3	PLP3	Skewness	O	-	CPP	95% percentile	O
3	HFCC	Maximum position	Δ	-	B4	Range	Δ
2	PLP2	Regression coefficient	O		SF slope	Minimum	O
5	PLP4	slope	O	10	PLP3	Range	O
Anger L X Happiness H				11	HFBS	80% percentile	Δ
-	B4	range	O	13	MFCC	20% percentile	O
11	HFCC	skewness	O	13	MFCC	20% percentile	O
8	LFCC	10% percentile	Δ	7	MELBS	Maximum position	Δ
15	LFBS	Maximum	O	3	MFCC	Minimum	O
20	LFBS	Kurtosis	O	1	PLP3	Skewness	O
4	HFBS	Pearson skewness	Δ	8	B3	Range	Δ
12	LFCC	99% percentile	Δ	**Anger H X Happiness H**			
-	ZCR	kurtosis	O	3	LFCC	Median	O
-	SF slope	skewness	O	0	F0	20% percentile	Δ
8	LFCC	70% percentile	O	5	PLP1	std	O
Anger L X Neutral				5	HFBS	Minimum	Δ
3	PLP4	95% percentile	O	2	HFCC	60% percentile	O
3	LFCC	10% percentile	O	10	PLP4	Slope	O
2	PLP2	range	O	1	LFBS	Pearson skewness	O
7	PLP3	Relative std	O	19	LFCC	5% percentile	O
7	PLP3	skewness	O	2	PLP2	Skewness	O
Anger L X Sadness				9	B4	40% percentile	O
7	MELBS	Slope	O	**Anger H X Neutral**			
20	MELBS	80% percentile	O	-	4HzME	80% percentile	O
6	PLP4	slope	O	-	CPP	95 percentile	O
9	MFCC	slope	Δ	11	MFCC	Minimum	O
8	HFBS	95% percentile	Δ	-	F0	70% percentile	Δ
-	MSME	90% percentile	O	12	LFCC	99% percentile	O
11	MFCC	20% percentile	O	17	4HzME	Relative std	O

Coef.	Feature	High-level feature	P	Coef.	Feature	High-level feature	P
Happiness L X Happiness H				**Happiness L X Sadness**			
18	HFCC	Slope	O	6	PLP4	Relative std	O
4	LFCC	99% percentile	O	11	MFCC	70% percentile	Δ
11	LFBS	Median	Δ	1	LFBS	kurtosis	O
12	HFCC	Slope	Δ	18	LFCC	skewness	O
-	SF slope	Kurtosis	O	2	HFCC	99% percentile	O
-	B2	Range	Δ	13	LFCC	Regression coefficient	Δ
-	SF kurtosis	95% percentile	O	-	F0	70% percentile	O
17	MELBS	6th moment	O	16	LFCC	5% percentile	Δ
-	SF slope	5th moment	O	4	LFCC	Regression coefficient	Δ
2	LFCC	Relative minimum	O	**Happiness H X Sadness**			
Happiness L X Neutral				14	MELBS	90% percentile	Δ
8	HFCC	95% percentile	O	14	MFCC	10% percentile	Δ
1	LFBS	Pearson skewness	O	9	MFCC	Slope	O
1	PLP2	Relative maximum	O	**Neutral X Sadness**			
-	B4	5% percentile	O	2	HFCC	80% percentile	O
4	PLP3	skewness	O	1	MFCC	Skewness	O
3	HFCC	Maximum position	Δ	2	MFCC	Maximum	O
3	MELBS	Pearson skewness	O	9	LFBS	Minimum	Δ
3	LFCC	10% percentile	O	5	LFCC	Slope	O
9	LFCC	Maximum position	Δ	1	LFCC	80% percentile	O
3	PLP2	Median	O	-	ZCR	10% percentile	O
0	SF spread	10% percentile	O	15	LFBS	70% percentile	Δ
5	LFCC	Slope	O	17	HFBS	70% percentile	O
13	MFCC	10% percentile	Δ				
17	HFBS	70% percentile	O				
11	HFCC	Relative std	O				

7.3.10 Comparison of own system with relevant systems in literature

I believe that it is not possible to make a fair comparison of systems for emotion recognition when different speech databases are used. This is due to the variability in the number of emotions and their type, number of speakers, the matter of speech (clean/noisy) and the authenticity of emotional content. Despite of this, a brief comparison of own system with systems proposed in literature for emotion recognition in the domain of call centers are reported in Table 7.37.

Table 7.38: Comparison of own system with systems reported in literature.

System	Emotions	Feature extraction	Feature reduction	Classification	Best result
[YP07]	(2) Neutral and anger	Mean and standard deviation from F0, TE,MFCC	SFFS	Straightforward by *k*-NN or SVM	86.5% by SVM
[Yac+03]	(2) Neural and anger	Several High-level from F0, TE, and temporal features	FS	Straightforward by *k*-NN, ANN, SVM and DT	91% by SVM
[VS11]	(4) Neutral, nervous, querulous and other.	Several High-level from F0, MFCC, TE, SF	None	SVM	54%
[Lau+11]	(3) Irritation, Resignation and Neutral	F0, formants, TE, speech rate	Brutal force and FS	LDA	61.3%
[CD11]	(2) anger and neutral	F0, ZCR, MFCC, TE	None	SVM	85.3%
Own system	(7) see Table 7.16	See Table	mRMR+SFFS	Complex architecture	74.16% and 89.25% for surprise

The comparison reported in previous Table reveals that high classification accuracy of spontaneous emotions in the domain of call centers are performed with high accuracy when only two emotional states (anger and neutral) are considered. However, own approach works with 6 different emotional states. In case that only two states namely high anger and neutral are taken into account from the six reported in Table 7.31, then the proposed system outperforms approaches presented in [Yac+03, CD11, YP07], giving classification accuracy of 97.5%. The systems presented in [Lau+11, VS11] works with 3 and 4 emotional states respectively. Again, the proposed system significantly outperforms these mentioned systems.

7.3.11 Discussion

The results of straightforward classification presented by the usage of general classifier were not satisfactory. The best classification accuracy was achieved by combining segmental features reduced by applying TMV and suprasegmental features selected by mRMR and SFFS algorithm and SVM-RBF as a classifier, the classification accuracy for six emotional states under examination was 63.178%. It is also worth mentioning that the combination of both types of features improved the classification accuracy only by less than 1%. Considering the mentioned facts, it was necessary to find more sophisticated system to improve the performance. Hence, gender-dependent approach was tested; this approach showed improvement with classification accuracy of 67.86% for male speakers and 59.90% for female speakers.

Emotion coupling approach showed an excellent performance in terms of vocal emotion recognition from acted speech, thus it was tested on spontaneous speech as well. Results presented in section 7.3.5 suggested that emotion coupling approach can improve the classification accuracy by 5% comparing to general classifier. The best result was achieved for SVM-RBF as a classification element.

The next step was to fuse all systems mentioned above using one layer perceptron with 50 neurons. The fusion of system ensemble resulted in a significant improvement in terms of classification accuracy, giving an average result of 74.16% for six emotional states. The surprise state was detected independently by applying SVM with linear kernel function. This approach achieved F-measure of 89.25%.

The results of fusing network can be mapped into continuous two-dimensional space of emotions. This operation was carried by two approaches: the first one was based on the skeleton method previously proposed in section 7.2 and the second one was based on ANN. The second approach gave better results in terms of MAE (Table 7.33). However, it was tested only on Czech utterances labeled with respect to the activation-valence protocol described in section 7.2.

The general, gender-dependent and emotion coupling systems were trained by using utterances of all languages available in MSDES. This represents a drawback since it is known that emotion expression is language dependent. However, it was not possible to eliminate this drawback by creating independently trained models for each language since the number of training patterns was not sufficient. Nevertheless, the tests of the proposed system on the commercial level showed satisfactory results for all languages. One of the possible solutions for this issue is to adapt the system on the level of two-dimensional mapping by using different skeleton models for each language or by training the ANN using a subset of emotional utterances labeled with respect to the activation-valence protocol.

8 The influence of speaker's emotional state on gender recognition

As it was reported in chapter 7, emotion recognition accuracy can be improved when the classification scheme is gender-dependent. However, the gender classification itself might be influenced by the emotional content of speech. Hence, this chapter is devoted to the analysis of emotional speech from speaker's gender point of view and also proposes an effective approach for gender recognition.

From the relevant literature, it is observed that speaker's gender recognition has been mostly addressed as a complementary task to another speech processing applications. For example, it is proved that applying gender dependent approach might improve the speech or emotion recognition performance. However, the gender recognition itself is crucial for many applications, such as media retrieval or telephone call analysis.

The classification accuracy of speaker's gender depends upon some factors, like the speech quality, the number and length of speech records used for training/classification, speaker's age and speaker's emotional state. Hence, it is quite hard to make a fair comparison of algorithms reported in literature.

As it was mentioned above, one of the most important factors that might affect the gender recognition performance is the speaker's emotional state. Emotions with high activation and high valence are characterized by high levels of prosodic features, such as fundamental frequency and energy. Thus the speaker's gender can be confused due to the extreme variation of some features. This chapter is thus devoted to closely study the effect of this factor on the gender recognition process using MSDES database.

8.1 Experiments with neutral speech

This section section is devoted to tests made on neutral speech, which is the general case when gender recognition task is presented in literature [KVK11, YLQ08, Ich+10]. All features considered were tested separately and then, the best combination of them was found based on trial-error process. The classification accuracies of each feature and the best set are reported in Table 8.1. The classification accuracies were obtained for 30 high-level characteristics. These characteristics were identified by applying mRMR algorithm on the entire set of extracted features. The results suggested, as it appears from Table 8.1. That the combination of HFCC, F0 and formants based features yielded the best classification performance. Figure 8.1 illustrates the classification accuracy with respect to the number high-level features. This figure shows that by increasing the number of features no significant improvement of classification accuracy was observed. Hence, 30 features were used as a compromise between the classifier complexity and accuracy.

Table 8.1: Gender recognition results for different type of features.

Feature	Classification accuracy (%)
MFCC	92.7928
HFCC	94.054
LFCC	90.3495
PLP	92.7971
LPCC	83.9655
F0+Formants	95.3662
HFCC+F0+formants	96.1471

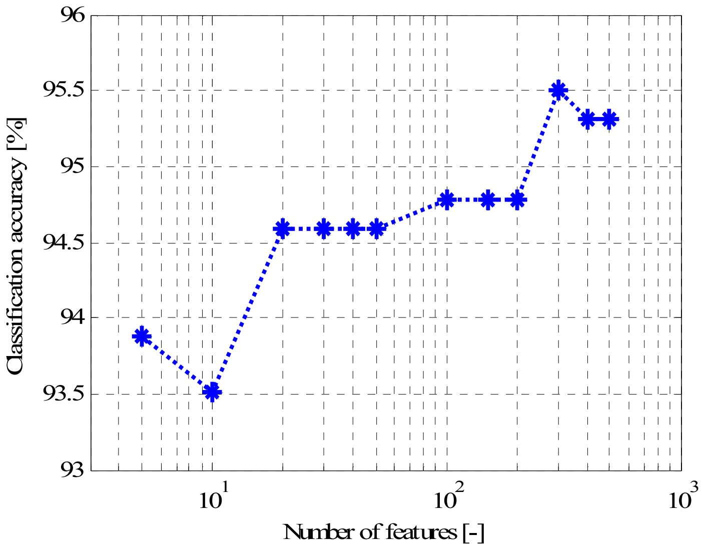


Figure 8.1: Classification accuracy against the number of selected high-level features.

8.2 Experiments with Emotional speech

In this section, I will deal with the emotional speech available from MSDES database described in chapter 7, taking into account only three emotional states: anger, happiness and sadness. Table 8.2 reports the classification accuracies of speaker's gender, when neutral speech is used for training. The features set here is that identified in previous experiment as optimal for gender recognition from neutral speech.

Results in Table 8.2 shows that the accuracies for all emotions are significantly lower in comparison with first experiment. This fact suggests that the usage of a classifier trained only with neutral speech can cause a high misclassification when emotional speech is considered. To avoid this, there are two ways to approach the task. The first one is to use a training database that contains emotional utterances, the second approach is more complicated and will be described later.

Table 8.2: Gender recognition results for different emotions when the classifier is trained by using neutral speech.

Emotion	**Classification accuracy (%)**
Anger	76.774
Happiness	87.078
Sadness	86.770

Considering the first approach, a new dataset is made consisting of mixed utterances for all states under examination (anger, happiness, sadness and neutral). Again, the optimal features for this new set were identified by applying mRMR algorithm. The classification accuracy of speaker's gender for this setup was 90.3%.

On the other hand, the second approach aims to find the best features for each emotion separately. This task is done by applying a modified Forward Selection (FS) algorithm, where in each iteration, only the feature that yields highest classification accuracy is selected. This algorithm is described in the following pseudocode.

```
SET $Z_0 = \emptyset$          //output feature group
SET $m = 1$           //feature index
SET $J(0) = 0$        //initialization of Classification accuracy function
SET $N_f = 2958$      //number of features considered
SET $N_{fs} = 50$     //number of FS iterations
```

```
FOR $m$=1 TO $N_{\mathrm{fs}}$    //the main cycle of FS algorithm
    FOR $n$=1 TO $N_{\mathrm{f}}$    // adding feature
        $D = T(\mathbf{F}_{\mathrm{neut}}(Z_{\boldsymbol{m}} \cup n))$  //Create a classifier D trained using neutral feature instances $\mathbf{F}_{\mathrm{neut}}$
        //Classify feature vector $\mathbf{F}^{i}_{\mathrm{emot}}$ by classifier D for all instances c and return the classification accuracy
        $y_c = C(\mathbf{F}_{\mathrm{emot}}(Z_{\boldsymbol{m}} \cup n), D)$
        $f^{+} = \mathrm{argmax}\ y_c$  //Find feature with highest accuracy
        $Z_{\boldsymbol{m+1}} = Z \cup f^{+}$  // add the selected feature to group Z
        $J(m) = \max(\ y_c)$    // update J
    ENDFOR   // end of feature adding cycle
ENDFOR   // end the main cycle of FS algorithm
```

Figure 8.2 illustrates the classification accuracy of speaker's gender with respect to the number of iterations of FS algorithm.

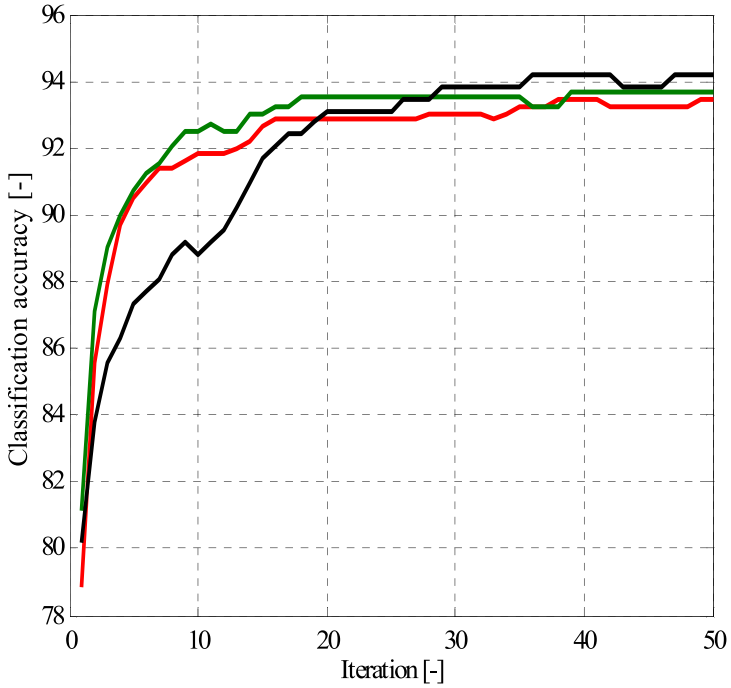


Figure 8.2: Classification accuracy of speaker's gender recognition against the number of forward selection iterations (black: sadness, red: anger and green: happiness).

The final set of features was slightly adjusted according to the classification accuracy in each iteration. For example, even though the highest accuracy for sadness was achieved for 50 features, only 29 of them were considered, as the difference in terms of accuracy is not significant. This step was taken in order to avoid any possible overfitting. The classification accuracies for different emotions obtained by applying the approach mentioned above are reported in Table 6.

Table 8.3: Gender recognition results after applying forward selection algorithm.

Emotion	**Classification accuracy (%)**
Anger	93
Happiness	93.50
Sadness	93.85

8.3 Summary

Gender recognition from spontaneous speech is performed with high accuracy (~96%) in case that only neutral speech is taken into account. The problem occurs when emotional speech is considered. Three scenarios were put under examination in this chapter. The first one showed as we mentioned above, the easiness of gender recognition from neutral speech. In the second scenario, the classifier trained using neutral speech was fed by emotional utterances; the accuracies for all emotions are significantly lower in comparison with first scenario, especially for anger, where the decrease was 20%. The third scenario was based on constructing one database involving emotional and non-emotional speech. The results of this setup showed an improvement comparing to previous scenario giving classification accuracy of 90.3%. The last scenario aimed at identifying the best features for each emotion separately by applying a modified forward selection algorithm. The outcome results from this approach were the best among scenarios that involve emotional speech (about 93.5%). In the light of all mentioned above, it is strongly recommended to take the emotional content of speech into account for real-life applications, either by involving the emotional speech in the training process or by adjusting the feature set to each emotion separately.

9 Analysis of spoken dialogue

The aim of the research proposed in this chapter [AS14] is to investigate the possibility of using dialogue features obtained from agent-client conversations to automatically identify successful phone calls in call centers. This can be very handy to spot the unsuccessful sessions within the large database of recorded telephone calls and can help the supervisors of call centers to figure out mistakes made by their agents. The basic block scheme of the proposed approach is illustrated in Figure 9.1 and the next sections are devoted to presenting each block in details.

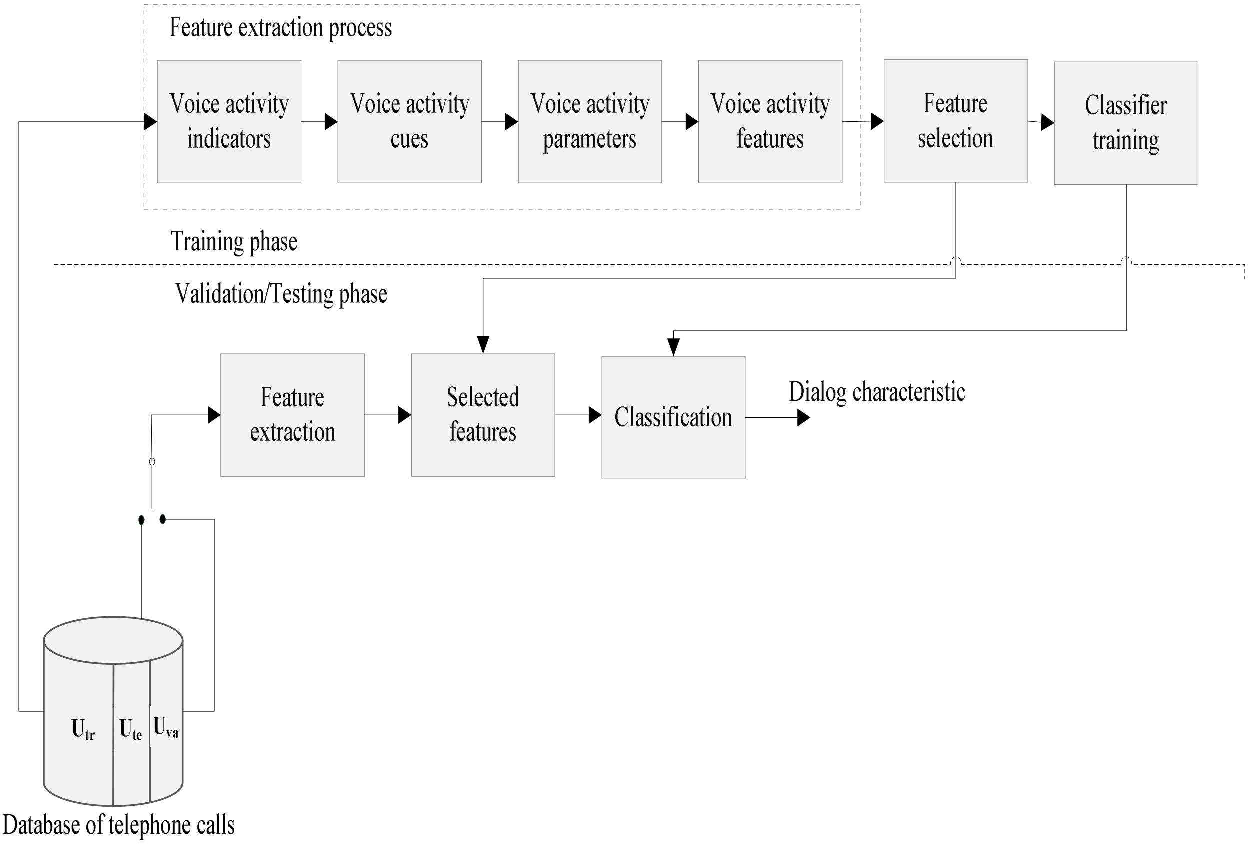


Figure 9.1: Basic block scheme of presented approach for successful call identification.

9.1 Description of speech corpus

The speech corpus used in experiments contains 48 dual-channel agent-client call records obtained from call centers in the Czech Republic. The basic characteristics of this corpus are reported in Table 9.1. It is worth mentioning that there is no statistical difference between the duration average values of successful and unsuccessful calls (p=0.29).

Table 9.1: Basic characteristics of speech corpus, the average duration of records and their standard deviation are in seconds.

Unsuccessful calls			**Successful calls**		
	duration			duration	
Count	average	std	Count	average	std
29	243	253	19	341	297

In spite of the easiness of determining which phone calls are successful or not even for a non-expert listener, a domain expert who worked as a supervisor in a Czech call center was asked to label the corpus. Phone calls from telemarketing domain were labeled as successful if the agent was able to sell a product to the client or convinced him to start using a certain service. The phone call was also considered successful when the agent was able to answer all clients' questions about a certain product or service which led the client to seriously think about buying the product or to start using the service in the future. Regarding calls from costumers' support services domain, the phone call was considered as successful when the agent was able to answer all costumers' questions and queries. All telephone records were then subjected to manual labeling in order to identify speech, silence and filled pauses periods. The outputs of the labeling process are called *voice activity indicators.*

9.2 Feature extraction

Four steps are defined within the process of feature extraction, these steps are

9.2.1 Extraction of voice activity indicators

In this step, the voice activity indicators are extracted either from the label file of each phone record or an automatic voice activity detector can be used to extract them. In our previous work [Mic+10], we proposed voice activity detection algorithm for the highly fluctuant recording conditions of call centers. The basic block scheme of this algorithm is illustrated in Figure 9.2.

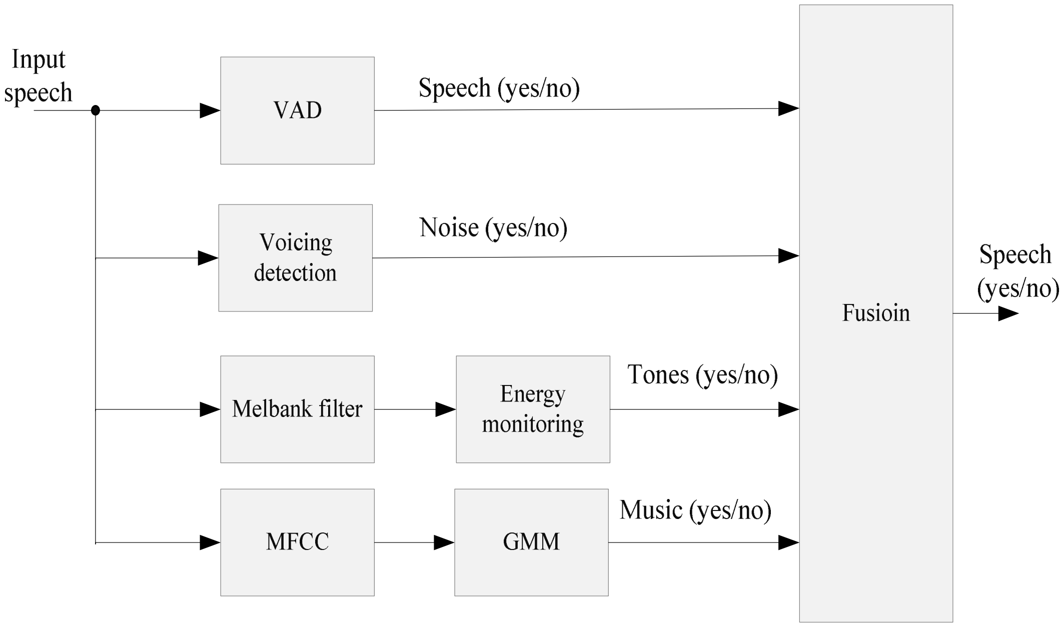


Figure 9.2: basic block scheme of voice activity detection algorithm for the highly fluctuant recording conditions of call centers.

9.2.2 Extraction of voice activity cues

This step involves the extraction of four different cues which are:

- **Hesitations**: silence periods or filled pauses that occur within the voice activity waveform in one direction that are not followed by voice activity in the other direction.
- **Reactions** (or turn takings): A reaction is registered when uninterrupted voice activity in one direction is followed by voice activity in the other direction.
- **Interruptions** (or overlaps): An interruption is registered when voice activity occurs simultaneously in both channels. The cues mentioned above are graphically illustrated on Figures 9.3 and 9.4.
- **Cumulative voice activity (CVA)**: this cue is proposed as a simple indicator of voice activity distribution over time. The formula that describes CVA is as follows

$$cva[n] = \sum_{i=0}^{n} \frac{vad[i]}{n}. \quad n = 0{,}1{,}2, \dots, N; \; i = \; 0{,}1{,}2 \dots, n. \tag{9.1}$$

Where $vad[i]$ is the i^{th} sample of the voice activity indicator with length of N.

9.2.3 Extraction of voice activity parameters

For all cues mentioned above except CVA, two vectors are constructed: one contains the position where the cue occurs and the second contains the length of the cue. For example, if three hesitations were registered in one direction, then the position vector and length vector might look like

H_p = (64000, 328000, 562000),

H_L = (4000, 8000, 2500).

For sampling frequency 8 kHz, it means the first hesitation occurred at sample 64000 (8th second) with length of 4000 samples (0.5 seconds).

9.2.4 Extraction of voice activity features

Statistical features are computed from vectors extracted in previous step and from CVA as well, these features are subsequently concatenated into the final feature vector used for training. The list of statistical features is reported in Table 9.2.

The result of feature extraction process is a vector of 490 statistical features extracted from both channels. For better understanding, the numbers of statistical features extracted from each cue/parameter are given in Table 9.3.

After obtaining voice activity features, SFFS algorithm was exploited in order to identify features that showed the maximum capability of discriminating between features representing successful and unsuccessful phone calls.

Table 9.2: List of statistical features.

Basic characteristics	mean, median, standard deviation, maximum, minimum, range, slope
Positional characteristics	position of maximum, position of minimum
Relative characteristics	relative standard deviation, relative range, relative maximum, relative minimum, relative position of maximum, relative position of minimum
moments	kurtosis, skewness, Pearson's skewness coefficient, 5th moment, 6th moment
Regression characteristics	linear regression coefficient, linear regression error
percentiles	1%, 5%, 10%, 20%, 30%, 40%, 60%, 70%, 80%, 90%, 95% and 99% percentile

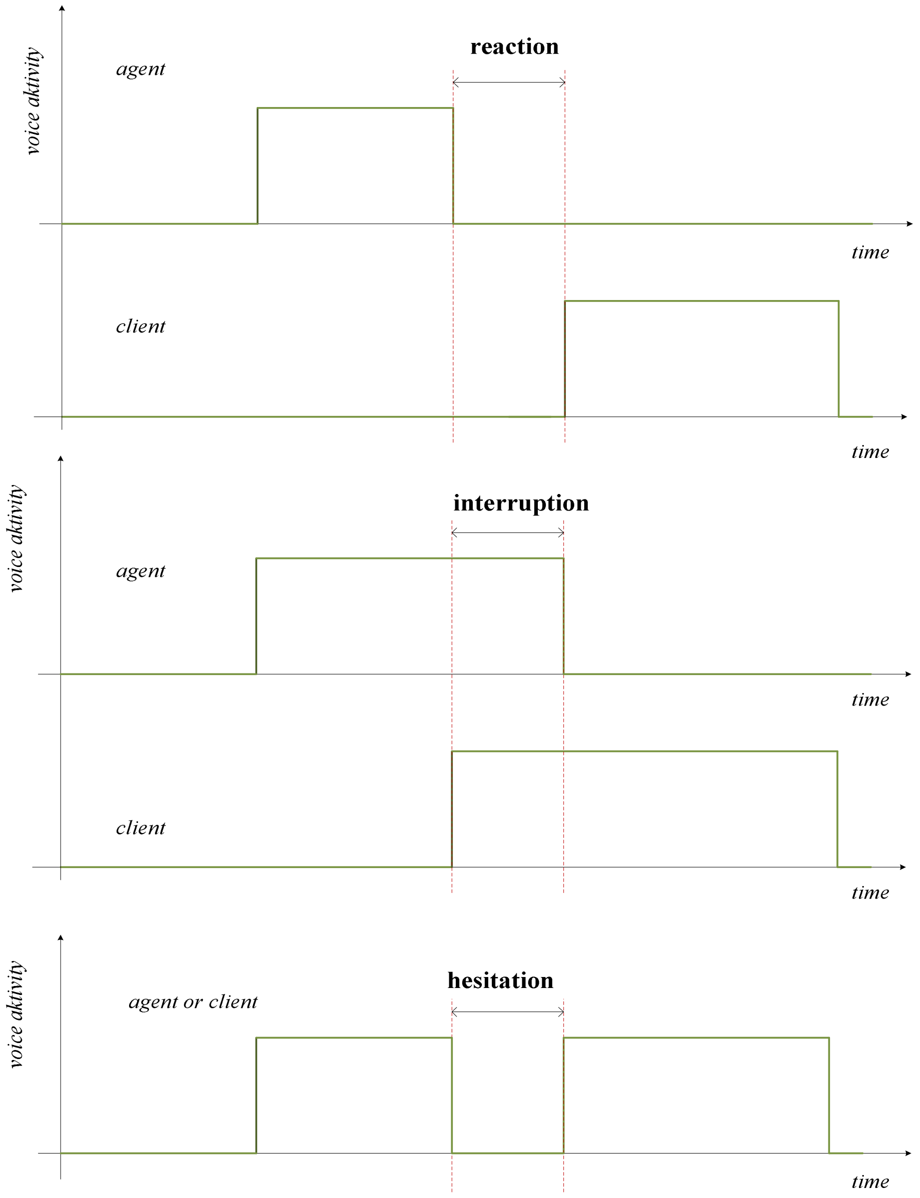


Figure9.3: Illustration of hesitation, reaction and interruption cues.

Table 9.3: The number of statistical features extracted from each cue/parameter.

Cue	**Parameter**	**No. statistical features**
Hesitations	Position vector	35
	Length vector	35
Interruptions	Position vector	35
	Length vector	35
Reactions	Position vector	35
	Length vector	35
CVA	-	35
Total for one channel		245
Total for two channels		**490**

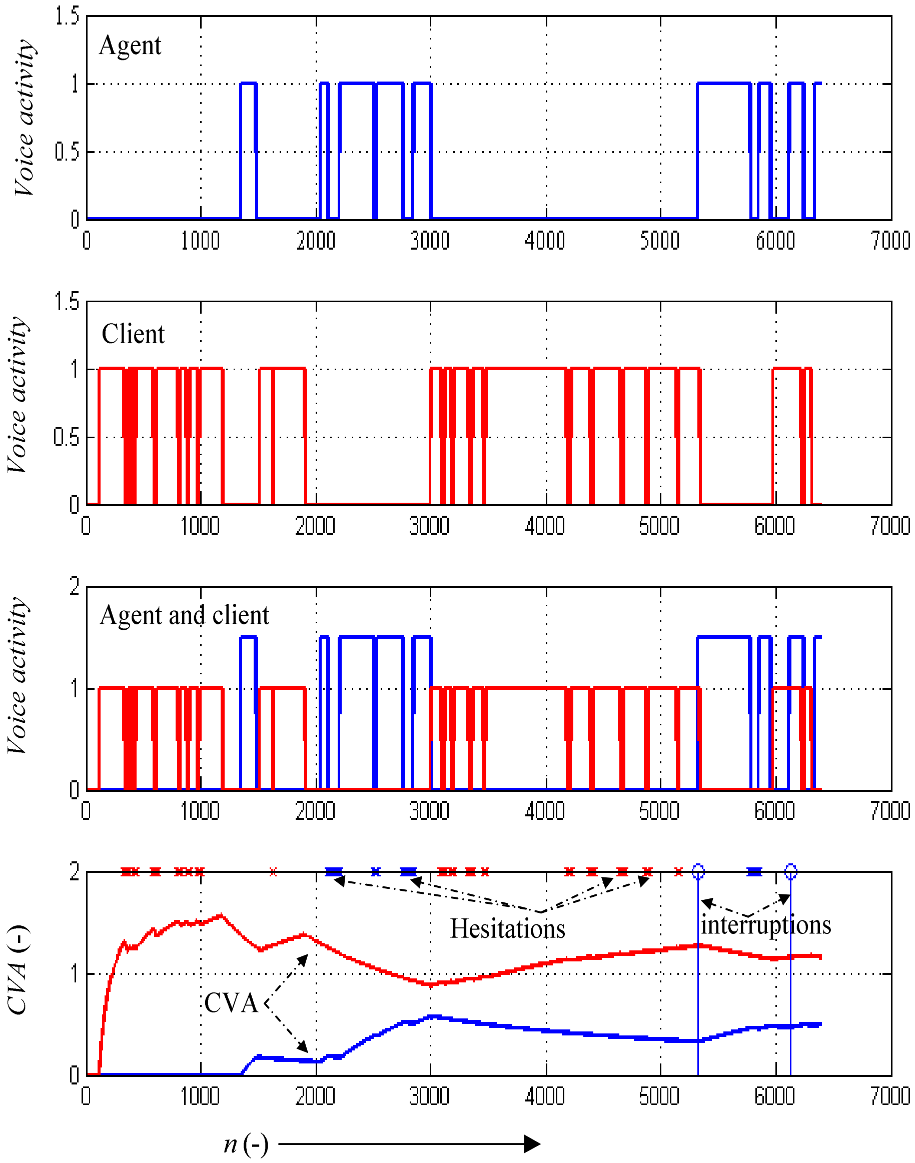


Figure9.4: Illustration of agent's and client's voice activity indicators (up) and the corresponding cumulative voice activity cues (down).

9.3 Experiment results

After extracting features and identifying those with most discriminative power for each classifier, leave-one-out validation is performed in order to assess the performance of these classifiers. Because our two classes (successful and unsuccessful phone calls) are not equally distributed, different evaluation measurements are reported, namely weighted accuracy, unweighted accuracy, precision, recall, F-measure and Matthews correlation. Moreover, ROC curves of each classifier are shown in Figure 9.5.

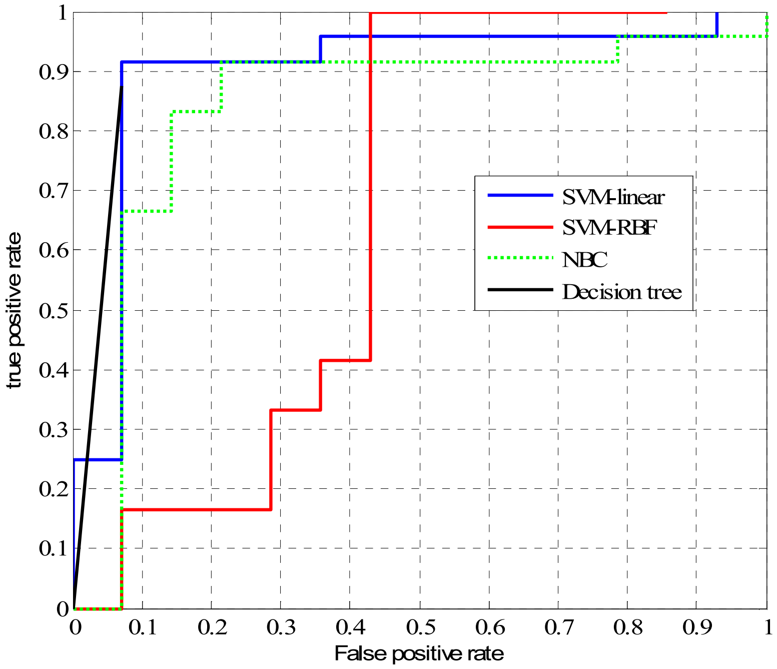


Figure 9.5: ROC curves of each classifier.

The experiment results reported in Table 9.4 shows that SVM classifier with linear function performed the best in terms of precision. However, this classifier evinced low recall comparing to NBC and SVM-RBF. The last mentioned classifier had, beside NBC, perfect (100%) recall rate. Nonetheless, SVM-RBF showed the lowest precision among all classifiers. Since both recall and precision are considered as important factors, the F-measure was selected as a criterion for classifier selection as this measurement takes into accounts both recall and precision. In terms of this criterion, NBC classifier performed the best with F-measure of 96%.

Table 9.5 contains features selected using SFFS for NBC. These features are reported in descending order from the most relevant to the least relevant. It is worth mentioning here that first four features were selected within SFFS as most relevant for all classifiers under examination.

Table 9.4: Experiment results in terms of different evaluation measurements.

	SVM-RBF	**SVM-linear**	**NBC**	**DT**
Weighted accuracy	84.21%	92.10%	94.74%	89.47%
Unweighted accuracy	78.58%	92.26%	92.86%	90.18%
Precision	80%	95.65%	92.31%	95.45%
Recall	100%	91.67%	100%	87.50%
F-measure	88.89%	93.62%	96%	91.30%
Matthews correlation	0.68	0.834	0.8895	0.79

Table 9.5: Most significant dialogue features.

Cue	**Direction**	**Parameter**	**Feature**
CVA	Client	-	minimum
Reaction	Agent	Position	minimum
Interruption	Agent	Length	10% percentile
CVA	Agent	-	Standard deviation
Hesitation	Client	Length	20% percentile
Hesitation	Client	Position	Relative maximum
Interruption	Client	Position	Slope
Reaction	Client	Length	Minimum
Hesitation	Client	Position	20% percentile

Further investigation of selected features revealed some interesting findings such as:

- For successful phone calls, the average minimum value of clients' CVA is approximately 5 times higher comparing to unsuccessful phone calls ($p<0.01$). This feature indicates that in successful calls the clients are more active in terms of voice activity.
- The second most relevant feature for detecting successful calls is the minimum position of Agents' reaction. This feature simply refers to the time needed by the agents to make their first reaction within the phone call. The results showed that this feature is about 3 times higher for unsuccessful phone calls comparing to successful ones ($p<0.01$).
- The agent's variability of CVA represented by the standard deviation is about 2.5 times higher for successful calls ($p<0.05$).
- The minimum length of clients' reaction in successful calls is about 2 times higher comparing to unsuccessful calls ($p<0.05$).

10 Conclusions

Emotion recognition from speech was a hot topic for researchers in the last decade. Hundreds of papers have been published so far on this topic proposing different approaches for vocal emotion recognition. Most of these papers presented results of experiments on acted speech, only a small fraction of the research was devoted to emotion recognition from spontaneous speech and very little attention has been given to emotion recognition from phone calls in call centers.

Regarding emotion recognition from acted speech, a new speaker independent procedure for classifying vocal expressions from Berlin Database of Emotional speech was proposed. The procedure is based on the splitting up of the emotion recognition process into two steps. In the first step, a combination of selected acoustic features is used to classify six emotions through a Bayesian Gaussian Mixture Model classifier (GMM). The two emotions that obtained the highest likelihood scores are selected for further processing in order to discriminate between them. For this purpose, a unique set of high-level acoustic features was identified using the Sequential Floating Forward Selection (SFFS) algorithm, and a GMM was used to separate between each couple of emotion. The mean classification rate is 81% with an improvement of 6% with respect to the most recent results obtained on the same database (75%).

Another new speaker-independent approach was introduced to the classification of emotional vocal expressions by using the COST 2102 Italian database of emotional speech which contains utterances of 6 basic emotional states: happiness, sarcasm/irony, fear, anger, surprise, and sadness. The proposed system was able to classify the emotions under examination with 60.7% accuracy by using a hierarchical structure consisting of a Perceptron and fifteen GMM trained to distinguish within each pair (couple) of emotions under examination. The best features in terms of high discriminative power were identified among a large number of spectral, prosodic and voice quality features. The results were compared with the subjective evaluation of the stimuli provided by human subjects.

The high-level features showed excellent discriminative power in terms of distinguishing between emotional states and therefore one section of this thesis was devoted to the analysis of these features. Results showed that the best high-level features in terms of high discriminative power strongly differ among the databases considered on the first hand and among the emotions within each database on the second hand. The second part of this thesis was devoted to emotion recognition using multilingual databases of spontaneous emotional speech, which is based on telephone records obtained from real call centers. The knowledge gained from experiments with emotion recognition from acted speech was exploited to design a new approach for classifying seven emotional states and mapping emotions into two dimensional space. The core of the proposed approach is complex classification architecture based on the fusion of different systems namely general, emotion coupling and gender dependent. The fusion of all systems achieved classification accuracy of 74.16% for six emotional states: High level anger, low level anger, high level happiness, neutral and sadness. The surprise state was detected separately because speakers can express surprise in both positive and negative way. The detection of surprise was carried out with high accuracy showing F-measure of 89.25% by using SVM with linear kernel.

The proposed emotion recognition engine is implemented in commercial system of Retia Company called ReDat [Ret], which proves its usability in real-life applications. Moreover, to our best knowledge, it is the first emotion recognition system exploited commercially for emotional analysis of calls in call centers. The system was also patented.

Finally, a new method was proposed to automatic identification of successful phone calls in call centers exploiting dialogue features. This approach can be very useful to spot the unsuccessful sessions within the large database of recorded telephone calls and can help the supervisors of call centers to figure out mistakes made by their agents. The features used for decision making are extracted from four cues namely hesitation, reaction, interruption and cumulative voice activity. The results achieved suggested that these features have a strong discriminative power in terms of classification between successful and unsuccessful phone calls showing F-measure of 96% by using Naïve Bayesian Classifier. All experiments presented in this thesis were carried out by using own tool called Hila (Appendix A).

A Hila

Hila is a modular tool with a graphical interface written in Matlab mainly used for experiments with speech processing and pattern recognition. All experiments reported in this thesis were carried out by using this tool. Hila contains native functions from Matlab, functions from other authors and mainly functions developed by the author of this thesis. Hila is being used by other researchers and students at Brno University of Technology, Czech Republic and University of Stirling, UK.

Why Hila was developed? In spite of existence of several tools that can be used for experiments with pattern recognition on speech processing. These tools appeared to have some shortages, for example:

- **RapidMiner** [Mie+06]: a powerful tool for pattern recognition tasks which contains various algorithms for feature selection, validation, classification and regression. However, it doesn't provide any functions for feature extraction from speech.
- **Weka** [HFH09]: a very popular tool for pattern recognition tasks due to its simplicity. This tool is useful when structured data are available so no feature extraction is needed to be performed.
- **Praat** [Boe02]: a very well-known tool for speech analysis providing only FFBP-NN for pattern recognition. Moreover, it doesn't cover a wide range of speech features.
- **HTK Toolkit**: The best tool so far for experiments with speech recognition. However, it doesn't have any graphical interface and for more complex approaches, such as feature fusing or using complex classification schemes.

The next is devoted to brief description of modules implemented in Hila

A.1 Database management module

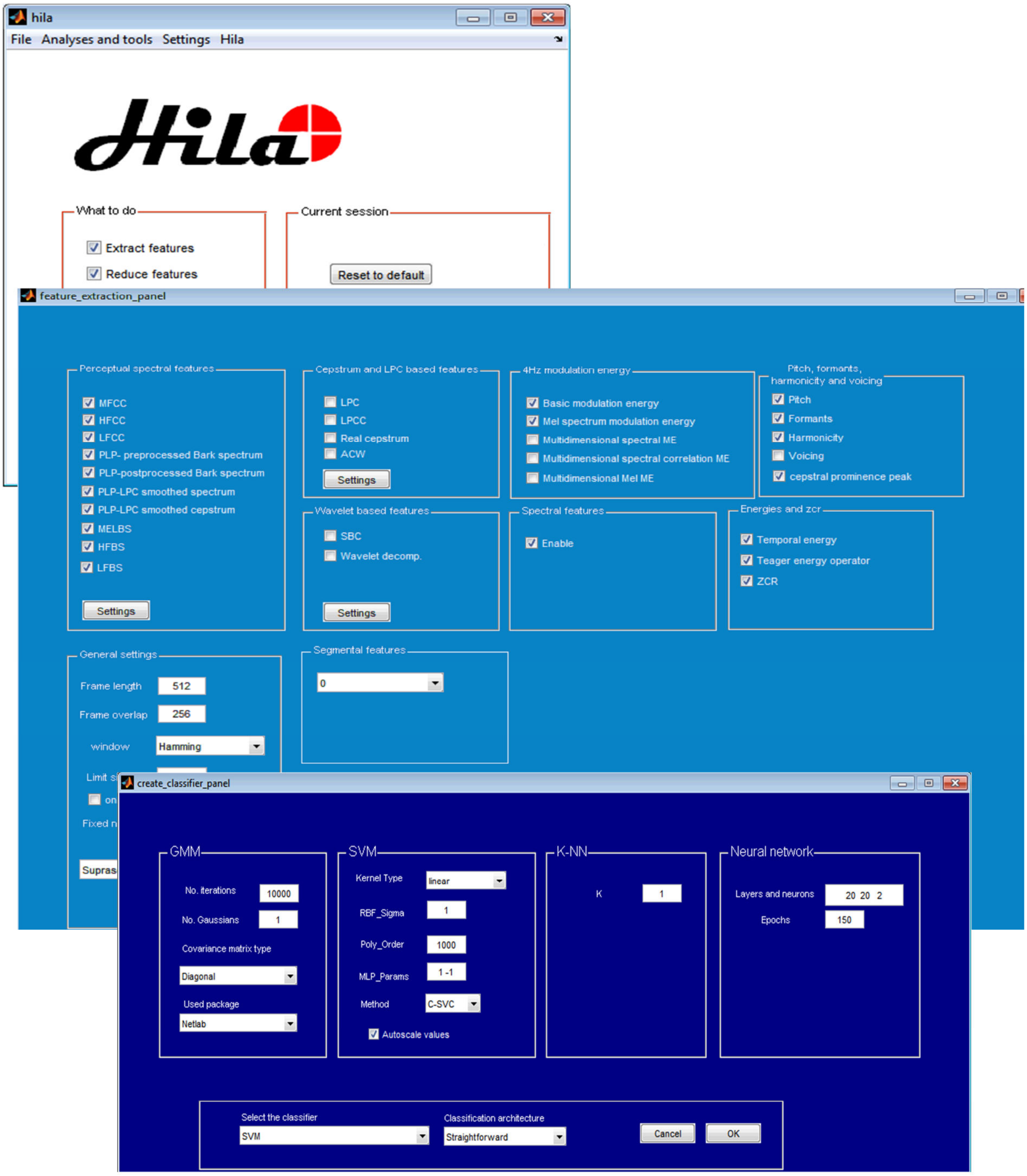
Defines the database path and both classes and speakers flags. Allow the user to define the training and testing set. The speaker dependency of the classification process is carried out in this module.

A.2 Preprocessing module

It allows the user to define operations applied on speech signal before passing it to the feature extraction module. This module contains the following functionalities

- **Resampling**: change the sampling frequency of input audio signal
- **Convert stereo to mono**: reduce the number of channels

- **Normalize**: Set the maximum level of input signal to 1
- **Filtering**: apply an arbitrary linear filter on input signal
- **Pre-emphasis**: applies Pre-emphasis HP filter on speech signal
- **Silence removing**: remove silence parts of speech



What to do
☑ Extract features
☑ Reduce features

Current session

Perceptual spectral features
☑ MFCC
☑ HFCC
☑ LFCC
☑ PLP- preprocessed Bark spectrum
☑ PLP-postprocessed Bark spectrum
☑ PLP-LPC smoothed spectrum
☑ PLP-LPC smoothed cepstrum
☑ MELBS
☑ HFBS
☑ LFBS

Cepstrum and LPC based features
□ LPC
□ LPCC
□ Real cepstrum
□ ACW

4Hz modulation energy
☑ Basic modulation energy
☑ Mel spectrum modulation energy
□ Multidimensional spectral ME
□ Multidimensional spectral correlation ME
□ Multidimensional Mel ME

Pitch, formants, harmonicity and voicing
☑ Pitch
☑ Formants
☑ Harmonicity
□ Voicing
☑ cepstral prominence peak

Wavelet based features
□ SBC
□ Wavelet decomp.

Spectral features
☑ Enable

Energies and zcr
☑ Temporal energy
☑ Teager energy operator
☑ ZCR

General settings
Frame length 512
Frame overlap 256
window Hamming
Limit si
□ on
Fixed n
Supras

Segmental features
0

GMM
No. iterations 10000
No. Gaussians 1
Covariance matrix type
Diagonal
Used package
Netlab

SVM
Kernel Type linear
RBF_Sigma 1
Poly_Order 1000
MLP_Params 1 -1
Method C-SVC
☑ Autoscale values

K-NN
K 1

Neural network
Layers and neurons 20 20 2
Epochs 150

Select the classifier SVM
Classification architecture Straightforward

Figure A.1: Screenshots of some GUI of Hila tool.

A.3 Feature extraction module

This module enables the user to choose and adjust the features extracted from speech signals. The following options can be set

- Segment length and segment overlap.
- Windowing function.
- Type of extracted features: segmental features, suprasegmental features or both.
- Extracted suprasegmental features.
- Number of differences extracted from each feature.

Table A.1 reports feature extraction techniques in Hila.

Table A.1: List of features in Hila.

Perceptual spectral features	**Spectral features**
Mel Frequency Cepstral Coefficients	Spectral centroid
Human Factor Cepstral Coefficients	Spectral spread
Linear Frequency Cepstral Coefficients	Spectral skewness
PLP preprocessed Bark Spectrum	Spectral kurtosis
PLP postprocessed Bark Spectrum	Spectral slope
PLP LPC smoothed spectrum	
PLP LPC smoothed cepstrum	
Mel Bank Spectrum	
Human Factor Bank Spectrum	
Linear Frequency Bank Spectrum	
Cepstrum and linear prediction based features	**Prosodic and voice quality features**
Linear Predictive Coefficients	Pitch
Linear Predictive Cepstral Coefficients	Formants
Real Cepstrum	Harmonicity
Adaptive Component Weighting	Voicing
	Cepstral prominence peak
	Temporal energy
	Teager energy operator
	Zero crossing ratio
Wavelet transform based features	
Suband Based Cepstral Coefficients (SBC)	

Wavelet decomposition based features	
Modulation energy based features	
Basic modulation energy Mel spectrum modulation energy Multidimensional spectral modulation energy Multidimensional spectral correlation modulation energy Multidimensional Mel modulation energy	

A.4 Validation module

Allows the user to choose among the following validation schemes

- **Hard assignment**: the training and testing patterns can be chosen manually.
- **Random cross validation**: the number of training and testing patterns and the number of validation cycles can be set by the user. The training and testing patterns are randomly selected
- **Leave one out validation**: The number of validation cycles is the same like the number of available patterns; this makes this method the most time consuming one. In each iteration, all patterns but one is used for training whereas the remaining pattern is used for testing.
- **Leave one speaker out validation**: this technique is very similar to the previous one. However, the number of validation cycles is defined by the number of speakers in the dataset. In each iteration, all patterns from one speaker are left for testing whereas the remaining patterns are used for training. This validation method ensure the speaker independency
- **N-Fold cross validation:** The dataset is equally split into *N* subgroups of patterns. In each iteration, all patterns from all subgroups but one are used for training, whereas the patterns of the remaining subgroup is used for testing.

A.5 Classification module

GMM: Apply Gaussian Mixture Model classifier with the following parameters

- Number of iterations of the EM algorithm
- Number of Gaussians in the mixture
- Type of covariance matrix: either diagonal or full

SVM: apply Support Vector Machine classifier

- Kernel type: Linear, polynomial, radial basis function, sigmoid

***k*-NN**: Apply *k*-Nearest Neighbor classifier

- *k*: number of k nearest neighbors included for decision

ANN: Apply Artificial Neural Network classifier

- Define number of layers and neurons in each layer
- Number of epochs

***k*-NN regression**: Apply k-NN algorithm for regression.

Decision tree for regression.

Decision tree for classification.

A.5 Fusion module

Defines and manage the fusing scheme on two possible levels: feature level and classifier level. If the fusion is on feature level, then the suprasegmental and segmental feature vectors are concatenated making the final feature vector. If the fusion is applied on the classification level, then one classifier should be selected for each type of feature vectors

- No fusion, use segmental features: doesn't apply any fusion on feature level, instead it uses only segmental features.
- No fusion, use suprasegmental features: doesn't apply any fusion on feature level, instead it uses only suprasegmental features.
- No fusion, use available features: doesn't apply any fusion on feature level, instead it uses available features.

A.6 Feature reduction module

Provide methods for reducing feature space dimensionality for both segmental and suprasegmental features.

Segmental features

- Frame by frame: classify each feature vector from each speech segment separately
- Frame by frame, overall evaluation:
- Frame under frame: concatenate feature vectors extracted from each segment into one feature vector.
- Mean vector: reduce the feature space dimensionality
- k-means: apply k-means algorithm

- PCA: apply principal component analysis
- GMM supervector: create GMM supervector from input feature space
- Self-organizing map

Suprasegmental features

- No reduction: doesn't apply any feature reduction technique on the input feature space
- Hard assignment: define the indices of features to be used
- F-rate: select features based on inter-class and intra-class distances
- mRMR: apply the minimal redundancy maximal relevance algorithm
- Feature selection based on *p*-value obtained from t-test
- Forward selection: apply forward selection
- Backward selection: apply backward selection
- SFFS: apply Sequential Floating Forward selection algorithm
- Correlation based selection of relevant features.

Beside the modules mentioned above, Hila contains functionalities for the evaluation of classifier performance, reporting and parallel computing. Figure A.1 contains screen shots of Hila graphical interface and figure A.2 illustrates the basic block scheme of this tool.

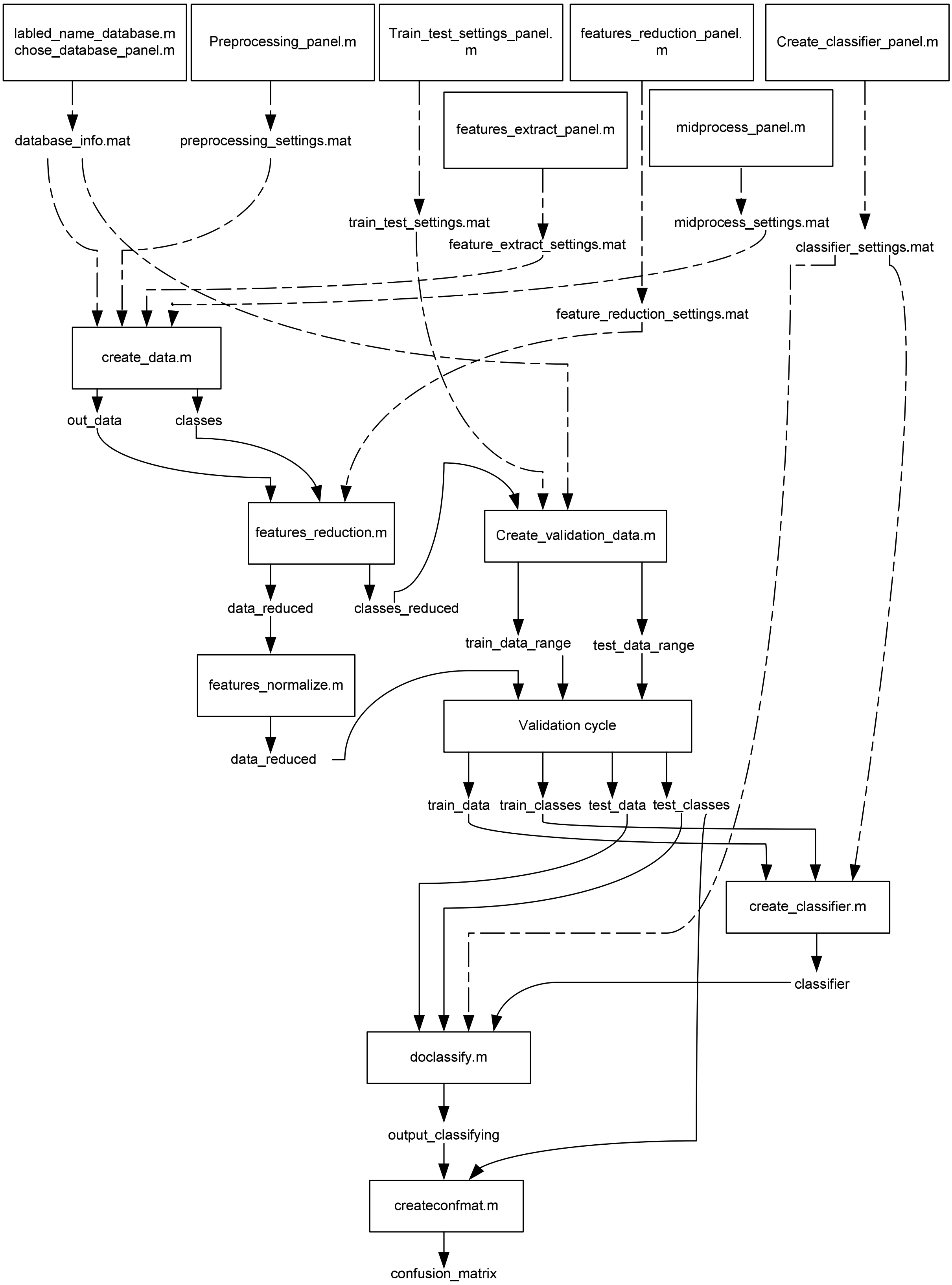
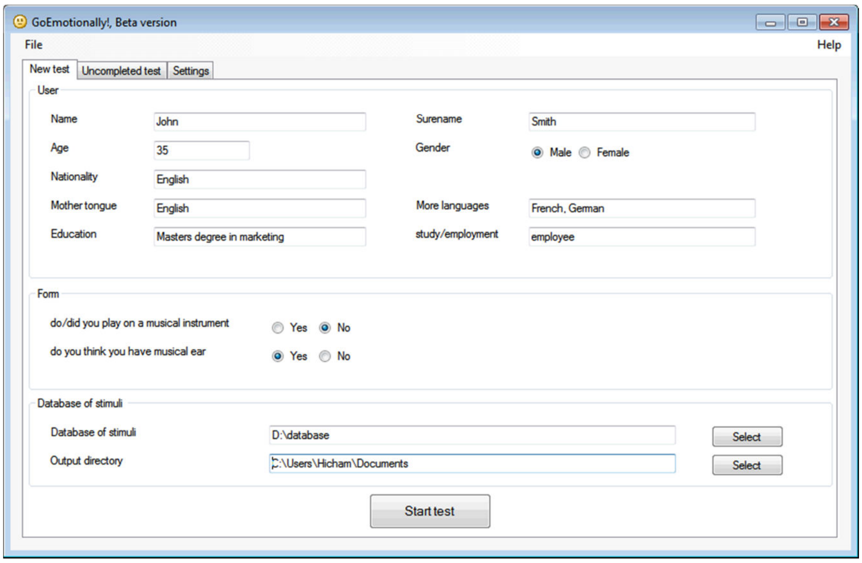


Figure A.2: Basic block scheme of Hila tool.

B GoEmotionally

GoEmotionally is a handy tool for subjective evaluation of multimedia content, which was primarily developed for subjective evaluation of humans emotional perception of multimodal material. It can be used for several purposes, such as subjective evaluation of denoising, compressing and TTS algorithms,



GoEmotionally!, Beta version

File Help

New test | Uncompleted test | Settings

User

Name: John

Surename: Smith

Age: 35

Gender: ● Male ○ Female

Nationality: English

Mother tongue: English

More languages: French, German

Education: Masters degree in marketing

study/employment: employee

Form

do/did you play on a musical instrument ○ Yes ● No

do you think you have musical ear ● Yes ○ No

Database of stimuli

Database of stimuli: D:\database Select

Output directory: C:\Users\Hicham\Documents Select

Start test

Figure B.1: GUI for submitting basic information about the subjective evaluator.

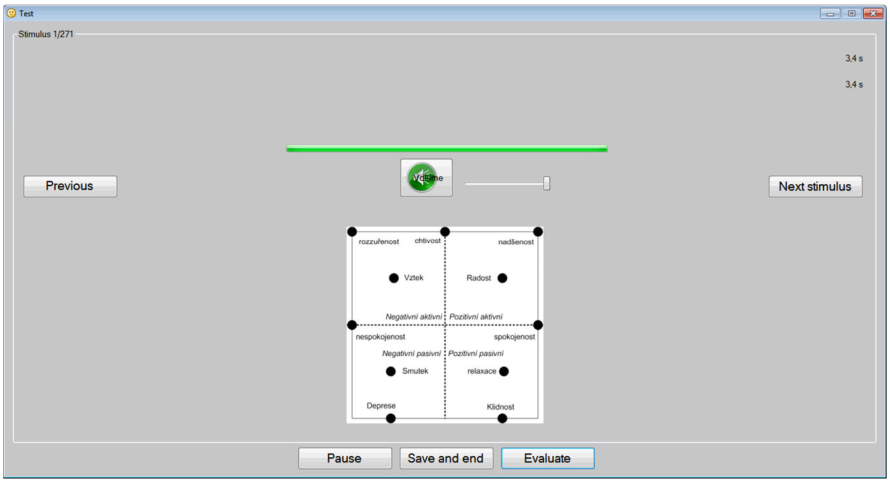


Figure B.2: Subjective evaluation screen for two-dimensional emotional space.

C Brief description of autonomous system for call centers surveillance and assessment

The system was developed within and industrial project with Czech company Retia “Advanced speech analyses technology for call centers and security services” financed by the Ministry of Industry and Trade of the Czech Republic. The aim of this project was to develop algorithms for emotion, age and gender recognition from telephone records in call centers.

Demands

- High reliability in terms of classification accuracy
- Low computational complexity
- Multilingual analysis (to deal with different languages)
- Real-time processing

Proposed system

- One-dimensional and two-dimensional interpretation of emotion recognition results.
- Voice activity detection and dialog analysis.
- Age and gender recognition.
- Multichannel real-time processing.
- 15x faster than real time.

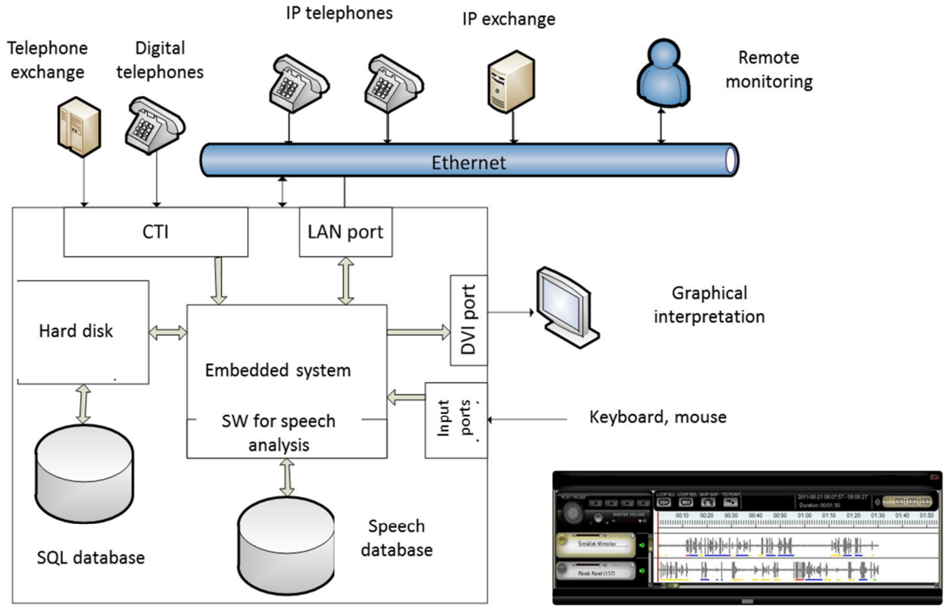


Figure C.1: Hardware block scheme of autonomous system for call centers surveillance and assessment.

The computational complexity of proposed system is summarized in Table C.1. The tests were carried out on a PC with Quad-core AMD Opteron 2.7 GHz processor and 8MB RAM on 64 bit windows server 2008 OS.

Table C.1: Results of computational complexity analysis for different module of proposed system for automatic analysis of phone call records.

File length [sec]	Only VAD		VAD + emotions		VAD + emotions + age recognition	
	Time [sec]	ratio	time [sec]	ratio	time [sec]	ratio
588	4.7	125	30.5	19	39	15
821	6.5	126	46	18	63	13
895	6.5	137	44	20	64	14

References

[AA00] Åsa Abelin and Jens Allwood. "Cross linguistic interpretation of emotional prosody." In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. 2000.

[AAE00] Bruno Apolloni, Guido Aversano and Anna Esposito. "Preprocessing and Classification of Emotional Features in Speech Sentences.". In *Proc. of International Workshop on Speech and Computer, Y. Kosarev (ed), SPIIRAS*, pp. 49-52 (2000)

[ACS09] Omar AlZoubi, Rafael A. Calvo, and Ronald H. Stevens. "Classification of EEG for affect recognition: an adaptive approach." In *AI 2009: Advances in Artificial Intelligence*, pp. 52-61. Springer Berlin Heidelberg, 2009.

[AE08] Hicham Atassi, and Anna Esposito. "A speaker independent approach to the classification of emotional vocal expressions." In *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, vol. 2, pp. 147-152. IEEE, 2008.

[AES11] Hicham Atassi, Anna Esposito, and Zdenek Smekal. "Analysis of high-level features for vocal emotion recognition." In *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*, pp. 361-366. IEEE, 2011.

[AHS11] Hicham Atassi, Amir Hussain and Zdenek Smekal. "Find My Emotion in the Space: A Novel Approach to Vocal Emotion Recognition". *In 6th International Conference on Teleinformatics*. 2011. pp. 230-235.

[AL05] Athanassios Avramidis, and Pierre L'Ecuyer. "Modeling and simulation of call centers." In *Simulation Conference, 2005 Proceedings of the Winter*, pp. 9-pp. IEEE, 2005.

[AS08] Hicham Atassi, and Zdenek Smekal. "Real-Time Model for Automatic Vocal Emotion Recognition". In *Proceedings of 31th International Conference on Telecommunications and Signal Processing - TSP 2008.* 2008. p. 90-95.

[AS14] Hicham Atassi, and Zdenek Smekal. Automatic Identification of Successful Phone Calls in Call Centers Based on Dialogue Analysis. "*Proceedings of 5rd IEEE International Conference on Cognitive Infocommunications* (submitted).

[ASE12] Hicham Atassi, Zdenek Smekal, and Anna Esposito. "Emotion Recognition from Spontaneous Slavic Speech". *In Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012).* 2012.

[Ata+10] Atassi, Hicham, Maria Teresa Riviello, Zdeněk Smékal, Amir Hussain, and Anna Esposito. "Emotional vocal expressions recognition using the COST 2102 Italian database of emotional speech." In *Development of multimodal interfaces: active listening and synchrony*, pp. 255-267. Springer Berlin Heidelberg, 2010.

[Ata08] Hicham Atassi. "Metody detekce základního tónu řeči". *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz)*, 2008, vol. 2008, no. 4, p. 1-17.

[AW13] Laksamon Archawaporn, and Waranyu Wongseree. "Erlang C model for evaluate incoming call uncertainty in automotive call centers.*" In Computer Science and Engineering Conference (ICSEC), 2013 International*, pp. 109-113. IEEE, 2013.

[BDY07] Bong-Wan, K., Dae-Lim, C., Yong-Ju, L.: Speech/Music Discrimination Using Mel-Cepstrum Modulation Energy, *Springer Berlin / Heidelber*g, pp. 406-414. ISSN 0302-9743, 2007.

[BGC06] Buscicchio, Cosimo A., Przemysław Górecki, and Laura Caponetti. "Speech emotion recognition using spiking neural networks." In *Foundations of Intelligent Systems*, pp. 38-46. Springer Berlin Heidelberg, 2006.

[BLN09] Busso, Carlos, Sungbok Lee, and Shrikanth Narayanan. "Analysis of emotionally salient aspects of fundamental frequency for emotion detection."*Audio, Speech, and Language Processing, IEEE Transactions on* 17, no. 4 (2009): 582-596.

[Boe02] Boersma, Paul. "Praat, a system for doing phonetics by computer." Glot international 5, no. 9/10 (2002): 341-345.

[Bur+05] Felix , Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In *Interspeech*, vol. 5, pp. 1517-1520. 2005.

[CD11] Clément Chastagnol, and Laurence Devillers. "Analysis of Anger across several agent-customer interactions in French call centers." In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4960-4963. IEEE, 2011.

[Cha+09] Aruna Chakraborty, Amit Konar, Uday Kumar Chakraborty, and Amita Chatterjee. "Emotion recognition from facial expressions and its control using fuzzy logic." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 39, no. 4 (2009): 726-743.

[CK11] Jangsik Cho, and Shohei Kato. "Detecting emotion from voice using selective Bayesian pairwise classifiers." In *Computers & Informatics (ISCI), 2011 IEEE Symposium on*, pp. 90-95. IEEE, 2011.

[CKC08] Ginevra Castellano, Loic Kessous, and George Caridakis. "Emotion recognition through multiple modalities: face, body gesture, speech." In *Affect and emotion in human-computer interaction*, pp. 92-103. Springer Berlin Heidelberg, 2008.

[Con13] Conley D. Quinn. "Simulating abandonment using Kaplan-Meier survival analysis in a shared billing and claims call center." *In Winter Simulation Conference*, pp. 1805-1817. 2013.

[Cou04] Mark Coulson. "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence." *Journal of nonverbal behavior* 28, no. 2 (2004): 117-139.

[CS07] Jarosław Cichosz, and K. Slot. "Emotion recognition in speech signal using emotion-extracting binary decision trees." *Proceedings of Affective Computing and Intelligent Interaction* (2007).

[CV95] Corinna Cortes, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.

[DHD12] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification.* John Wiley & Sons, 2012.

[Dou+00] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. "A new emotion database: considerations, sources and scope." In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.* 2000.

[EA05] Anna Esposito, and Guido Aversano. "Text independent methods for speech segmentation." In *Nonlinear Speech Modeling and Applications*, pp. 261-290. Springer Berlin Heidelberg, 2005.

[EGP10] Humberto Pérez Espinosa, J. O. Garcia, and Luis Villasenor Pineda. "Features selection for primitives estimation on emotional speech." In *Acoustics Speech* *and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5138-5141. IEEE, 2010.

[Ekm92] Paul Ekman. "Facial expressions of emotion: New findings, new questions."*Psychological science* 3, no. 1 (1992): 34-38.

[ERM09] Anna Esposito, Maria Teresa Reviello, and Giuseppe Di Maio. "The COST 2102 Italian audio and video emotional database." In *Neural Nets WIRN09: Proceedings of the 19th Italian Workshop on Neural Nets, Vietri Sul Mare, Salerno, Italy May 28-30 2009*, vol. 204, p. 51. IOS Press, 2009.

[ESS07] Anna Esposito, Vojtěch Stejskal, Zdeněk Smékal, and Nikolaos Bourbakis. "The significance of empty speech pauses: cognitive and algorithmic issues." In*Advances in Brain, Vision, and Artificial Intelligence*, pp. 542-554. Springer Berlin Heidelberg, 2007.

[Far+13] Kamran Farooq, Amir Hussain, Hicham Atassi, Stephen Leslie, Chris Eckl, Calum MacRae, and Warner Slack. "A novel clinical expert system for chest pain risk assessment." In *Advances in Brain Inspired Cognitive Systems*, pp. 296-307. Springer Berlin Heidelberg, 2013.

[GLC03] Hyoun-Joo Go, Keun-Chang Kwak, Dae-Jong Lee, and Myung-Geun Chun. "Emotion recognition from the facial image and speech signal." In *SICE 2003 Annual Conference*, vol. 3, pp. 2890-2895. IEEE, 2003.

[GS05] Alvin I. Goldman, and Chandra Sekhar Sripada. "Simulationist models of face-based emotion recognition." *Cognition* 94, no. 3 (2005): 193-213.

[Her90] Hynek Hermansky. "Perceptual linear predictive (PLP) analysis of speech." *the Journal of the Acoustical Society of America* 87, no. 4 (1990): 1738-1752.

[HH73] Tammo Houtgast, and Herman JM Steeneken. "The modulation transfer function in room acoustics as a predictor of speech intelligibility." *Acta Acustica united with Acustica* 28, no. 1 (1973): 66-73.

[HFH09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update."*ACM SIGKDD explorations newsletter* 11, no. 1 (2009): 10-18.

[Hua+01] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Foreword By-Reddy.*Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall PTR, 2001.

[HSG02] Yolanda D. Heman-Ackah, Deirdre D. Michael, and George S. Goding Jr. "The relationship between cepstral peak prominence and selected parameters of dysphonia." *Journal of Voice* 16, no. 1 (2002): 20-27.

[Ich+10] Masatsugu Ichino, Naohisa Komatsu, Wang Jian-Gang, and Yau Wei Yun. "Speaker gender recognition using score level fusion by AdaBoost." In *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*, pp. 648-653. IEEE, 2010.

[JM10] Liv Jing, and Guo Min. "Predicting Call Center Service Grade with Improved Neural Network Algorithm." In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on*, pp. 1-4. IEEE, 2010.

[Kim+10] Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. "Music emotion recognition: A state of the art review." In *Proc. ISMIR*, pp. 255-266. 2010.

[KL06] Yi-hao Kao, and Lin-shan Lee. "Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language." In *InterSpeech*. 2006.

[KL12] Seliz Gülsen Karadogan, and Jan Larsen. "Combining semantic and acoustic features for valence and arousal recognition in speech." In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*, pp. 1-6. IEEE, 2012.

[Koe+10] Sander Koelstra, Ashkan Yazdani, Mohammad Soleymani, Christian Mühl, Jong-Seok Lee, Anton Nijholt, Thierry Pun, Touradj Ebrahimi, and Ioannis Patras. "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos." In *Brain informatics*, pp. 89-100. Springer Berlin Heidelberg, 2010.

[Kun04] Ludmila I. Kuncheva. *Combining pattern classifiers: methods and algorithms.* John Wiley & Sons, 2004.

[KVK11] Marko Kos, Damjan Vlaj, and Z. Kacic. "Speaker's gender classification and segmentation using spectral and cepstral feature averaging." In *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*, pp. 1-4. IEEE, 2011.

[Lau+11] Petri Laukka, Daniel Neiberg, Mimmi Forsell, Inger Karlsson, and Kjell Elenius. "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation." *Computer Speech & Language*25, no. 1 (2011): 84-104.

[Lef+10] Iulia Lefter, Leon JM Rothkrantz, Pascal Wiggers, and David A. Van Leeuwen. "Emotion recognition from speech by combining databases and fusion of classifiers." In *Text, Speech and Dialogue*, pp. 353-360. Springer Berlin Heidelberg, 2010.

[LM98] Huan Liu, and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Springer, 1998.

[LN03] Chul Min Lee, and Shrikanth Narayanan. "Emotion recognition using a data-driven fuzzy inference system." In *INTERSPEECH*. 2003.

[LY07] Marko Lugger, and Bin Yang. "The relevance of voice quality features in speaker independent emotion recognition." In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV-17. IEEE, 2007.

[LZZ10] Xiang Li, Xin Li, Xiaoming Zheng, and Dexing Zhang. "EMD-TEO Based speech emotion recognition." In *Life System Modeling and Intelligent Computing*, pp. 180-189. Springer Berlin Heidelberg, 2010.

[Mic+10] Ivan Mica, Hicham Atassi, Jiri Prinosil, and Petr Novak. "Voice activity detection under the highly fluctuant recording conditions of call centres."*Proceedings of ECS'10/ECCTD'10/ECCOM'10/ECCS'10* (2010): 334-336.

[Mie+06] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. "Yale: Rapid prototyping for complex data mining tasks." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 935-940. ACM, 2006.

[MXR96] Mammone J., Xiaoyu Z., Ravi P.: Robust Speaker Recognition: a Feature-based Approach. Vol. 13. *IEEE Signal Processing Magazine*, 1996.

[MGS07] Mporas I., Ganchev T., Siafarikas, M., Fakotakis, N.: Comparison of Speech Features on the Speech Recognition Task. *Journal of Computer Science*, pp: 608-616, 2007.

[MN11] Emily Mower, and Shrikanth Narayanan. "A hierarchical static-dynamic framework for emotion classification." In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 2372-2375. IEEE, 2011.

[MR74] Albert Mehrabian , and James A. Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[MVG10] Kudiri K.Mohan, Gyanendra K. Verma, and Bakul Gohel. "Relative amplitude based features for emotion detection from speech." In *Signal and Image Processing (ICSIP), 2010 International Conference on*, pp. 301-304. IEEE, 2010.

[NHL06] Eva Navas, Inma Hernáez, and Iker Luengo. "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS." *Audio, Speech, and Language Processing, IEEE Transactions on* 14, no. 4 (2006): 1117-1127.

[NFT06] Norman D. Cook, Takashi X. Fujisawa, and Kazuaki Takami. "Evaluation of the affective valence of speech using pitch substructure." *Audio, Speech, and Language Processing, IEEE Transactions on* 14, no. 1 (2006): 142-151.

[NNT00] Ryohei Nakatsu, Joy Nicholson, and Naoko Tosa. "Emotion recognition and its application to computer agents with spontaneous interactive capabilities."*Knowledge-Based Systems* 13, no. 7 (2000): 497-504.

[Nwe+03] Tin Lay Nwe, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." *Speech communication* 41, no. 4 (2003): 603-623.

[OKJ06] Keith Oatley, Dacher Keltner, and Jennifer M. Jenkins. *Understanding emotions* . Blackwell publishing, 2006.

[Osg57] Charles Egerton Osgood,. *The measurement of meaning.* No. 47. University of Illinois press, 1957.

[Pao+07] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and Yu-Yuan Lin. "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech." In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, pp. 997-1005. Springer Berlin Heidelberg, 2007.

[PCY04] Tsang-Long Pao, Yu-Te Chen, and Jun-Heng Yeh. "Emotion recognition from mandarin speech signals." In *Chinese Spoken Language Processing, 2004 International Symposium on*, pp. 301-304. IEEE, 2004.

[PK12] Abhishek M. Pandharipande, and Sunil Kumar Kopparapu. "A novel approach to identify problematic call center conversations.*" In Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on,* pp. 1-

5. IEEE, 2012.

[PLD05] Hanchuan Peng, Fulmi Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*27, no. 8 (2005): 1226-1238.

[Pud+94] Pavel Pudil, F. J. Ferri, J. Novovicova, and J. Kittler. "Floating search methods for feature selection with nonmonotonic criterion functions." In *In Proceedings of the Twelveth International Conference on Pattern Recognition, IAPR*. 1994.

[PR11] Tomas Pfister, and Peter Robinson. "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis." *Affective Computing, IEEE Transactions on* 2, no. 2 (2011): 66-78.

[PS03] Timo Partala, and Veikko Surakka. "Pupil size variation as an indication of affective processing." *International journal of human-computer studies* 59, no. 1 (2003): 185-198.

[Pic00] Rosalind W. Picard. *Affective computing*. MIT press, 2000.

[PP06] Maja Pantic, and Ioannis Patras. "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences."*Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*36, no. 2 (2006): 433-449.

[Ret] http://www.redat.cz/cs/redat-voiceprocessor

[RLC09] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. "Acoustic feature selection for automatic emotion recognition from speech." *Information processing & management* 45, no. 3 (2009): 315-328.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing*10, no. 1 (2000): 19-41.

[SFP09] Stefan Scherer, Friedhelm Schwenker, and Günther Palm. "Classifier fusion for emotion recognition from speech." In *Advanced intelligent environments*, pp. 95-117. Springer US, 2009.

[SH00] Hansen R. Sarikaya, J.H.L: High Resoultion Speech Feature Parameterization for /monophone Based Stressed Speech Recognition. *IEEE Signal Processing Letters*, 2000.

[SH03] Mark D. Skowronski., and John G. Harris. "Improving the filter bank of a classic speech feature extraction algorithm." In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*, vol. 4, pp. IV-281. IEEE, 2003.

[SHC10] Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. "The OGI kids' speech corpus and recognizers." In *Proc. of ICSLP*, pp. 564-567. 2010.

[Sche00] Klaus R. Scherer. "Emotion effects on voice and speech: Paradigms and approaches to evaluation." In *Presentation held at ISCA Workshop on Speech and Emotion, Belfast*, vol. 10. 2000.

[Sch+08] Stefan Scherer, Mohamed Oubbati, Friedhelm Schwenker, and Günther Palm. "Real-time emotion recognition from speech using echo state networks." In*Artificial neural networks in pattern recognition*, pp. 205-216. Springer Berlin Heidelberg, 2008.

[Sun+11] Johan Sundberg, Sona Patel, Eva Bjorkner, and Klaus R. Scherer. "Interdependencies among voice source parameters in emotional speech."*Affective Computing, IEEE Transactions on* 2, no. 3 (2011): 162-174.

[SRL03] Björn Schuller, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2, pp. II-1. IEEE, 2003.

[SRL04] Björn Schuller, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. I-577. IEEE, 2004.

[SSP09] Stefan Scherer,, Friedhelm Schwenker, and Günther Palm. "Classifier fusion for emotion recognition from speech." In *Advanced intelligent environments*, pp. 95-117. Springer US, 2009.

[Ste+08] Stefan Steidl, Anton Batliner, Elmar Nöth, and Joachim Hornegger. "Quantification of segmentation and F0 errors and their effect on emotion recognition." In *Text, Speech and Dialogue*, pp. 525-534. Springer Berlin Heidelberg, 2008.

[Stu+11] André Stuhlsatz, Christine Meyer, Florian Eyben, T. ZieIke, Günter Meier, and Björn Schuller. "Deep neural networks for acoustic emotion recognition: raising the benchmarks." In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5688-5691. IEEE, 2011.

[SY12] Sivaraman Sriram, and Xiaobu Yuan. "An enhanced approach for classifying emotions using customized decision tree algorithm." In *Southeastcon, 2012 Proceedings of IEEE*, pp. 1-6. IEEE, 2012.

[TPH03] Min Tang, Bryan Pellom, and Kadri Hacioglu. "Call-type classification and unsupervised training for the call center domain." In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 204-208. IEEE, 2003.

[VA05] Thurid Vogt, and Elisabeth André. "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition." In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 474-477. IEEE, 2005.

[Vla11] Bogdan Vlasenko, David Philippou-Hubner, Dmytro Prylipko, Ronald Bock, Ingo Siegert, and Andreas Wendemuth. "Vowels formants analysis allows straightforward detection of high arousal emotions." In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pp. 1-6. IEEE, 2011.

[VK05] Dimitrios Ververidis, and Constantine Kotropoulos. "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm." In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 1500-1503. IEEE, 2005.

[VK06] Dimitrios Ververidis, and Constantine Kotropoulos. "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections." In *Proc. XIV European Signal Processing Conf.* 2006.

[VS11] Klára Vicsi, and Dávid Sztahó. "Problems of the Automatic Emotion Recognitions in Spontaneous Speech; An Example for the Recognition in a Dispatcher Center." In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pp. 331-339. Springer Berlin Heidelberg, 2011.

[Van84] Van Bezooijen. The Characteristics and Recognisability of Vocal Expression of Emotions. *Drodrecht, The Netherlands, Foris*, 1984.

[WF05] Ian H. Witten, and Eibe Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[WL11] Chung-Hsien Wu, and Wei-Bin Liang. "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels." *Affective Computing, IEEE Transactions on* 2, no. 1 (2011): 10-21.

[Yac+03] Sherif M. Yacoub, Steven J. Simske, Xiaofan Lin, and John Burns. "Recognition of emotions in interactive voice response systems." In*INTERSPEECH*. 2003.

[Yeg09] Yegnanarayana, B. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[Yeg78] Yegnanarayana, B. "Formant extraction from linear-prediction phase spectra."*The Journal of the Acoustical Society of America* 63, no. 5 (1978): 1638-1640.

[YLQ08] Fan Yingle, Yi Li, and Tong Qinye. "Speaker gender identification based on combining linear and nonlinear features." In *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pp. 6745-6749. IEEE, 2008.

[YP11] Won-Jung Yoon, and Kyu-Sik Park. "Building robust emotion recognition system on heterogeneous speech databases." *Consumer Electronics, IEEE Transactions on* 57, no. 2 (2011): 747-750.

[YP07] Won-Joong Yoon, and Kyu-Sik Park. "A study of emotion recognition and its applications." In *Modeling Decisions for Artificial Intelligence*, pp. 455-462. Springer Berlin Heidelberg, 2007.

[Yu+01] Feng Yu, Eric Chang, Yingqing Xu, and Heung-Yeung Shum. "Emotion detection from speech to enrich multimedia content." In *Proceedings of the second IEEE pacific rim conference on multimedia: Advances in multimedia information processing*, pp. 550-557. Springer-Verlag, 2001.

[ZHK01] Guojun Zhou, John HL Hansen, and James F. Kaiser. "Nonlinear feature based classification of speech under stress." *Speech and Audio Processing, IEEE Transactions on* 9, no. 3 (2001): 201-216.

[ZPR09] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*31, no. 1 (2009): 39-58.

Curriculum Vitae

Hicham Atassi

Personal information

Date of birth:	21th April, 1984
Nationality:	Czech
Email:	atassi@feec.vutbr.cz
Tel:	(+420) 776 560 452
Address:	Palackého třída 1623/6, Brno Czech Republic

Language Skills

Czech and Arabic:	Mother tongues
English:	Excellent skills (reading, writing, speaking)
Italian:	Basic conversation skills

Education

2007-2014 Brno University of Technology, Faculty of Electrical Engineering and Communication, Czech Republic, **doctoral degree** in Teleinformatics

2005-2007 Brno University of Technology, Faculty of Electrical Engineering and Communication, Czech Republic, **master's degree** in Telecommunication and Information Technology. ***Graduated with honors***. ***Awarded dean's prize for best master's thesis***

2002-2005 Brno University of Technology, Faculty of Electrical Engineering and Communication, Czech Republic, **bachelor's degree** in Teleinformatics

Professional Experience

2007-now Research and teaching assistant. Brno University of Technology, Department of Telecommunications, Czech Republic.

Teaching experience:

- Supervised more than 10 bachelor and master thesis

- Taught courses: Signal and System Analysis for undergraduates (practicals), Speech Processing for postgraduates (practicals and lectures)

2009-2010 Research and teaching assistant. University of Striling, Department of Mathematics and Computing Science, United Kingdom.

Teaching experience:

- Taught courses: Multimedia for postgraduates (practicals and lectures), Decision Support Systems for postgraduates (practicals), Programming and User Interface Design (practicals)

Research stays

Feb-May 2008 International Institute of Advanced Scientific Studies (IIASS). Salerno, Italy

June-July 2009 International Institute of Advanced Scientific Studies (IIASS). Salerno, Italy

May 2010-June 2010 University Pierre et Marie Curie, Institut des Systèmes Intelligents et de Robotique. Paris, France

Sep 2009 –May 2010 University of Stirling, Department of Mathematics and Computing Science, United Kingdom

membership

2008-2011 Member of the scientific committee of COST 2102 project “Cross Modal Analysis of Verbal and Nonverbal Communication”

2010- now Head of Human-Machine Interaction Group, Signal Processing Laboratory, Department of Telecommunications, Brno University of Technology

Invited talks

2013 “An Autonomous Intelligent System for Call-centers Monitoring and Assessment”. COST workshop on Social Robotics, Brussels, Belgium

2013 “Human-Machine interaction group of Signal Processing Laboratory”. Amity University, New Delhi, India

2010 “COST 2102 Action: Cross-Modal Analysis of Verbal and Non-verbal Communication”. Tishreen University, Syria

2008 “Emotion recognition from speech”. Audio Engineering Society, Prague, Czech Republic

Participation in projects

2011-2014 EE.2.3.20.0094 - Support for incorporating R&D teams in international cooperation in the area of image and audio signal processing. Funded by the Ministry of Education, Youth and Sports of the Czech Republic and European Union. Grantholders: Jan Karasek, Hicham Atassi et al.,

2010-2014 VG20102014033 - Improvement of risk area security using combined methods for biometrical identification of subjects. Funded by interior ministry of the Czech Republic. Grantholder: Kamil Vrba,

2009-2011 FR-TI1/481 - Advanced speech analyses technology for call centers and security services. Funded by the ministry of industry and trade of the Czech Republic. Grantholder: Kamil Vrba

2009-2010 Enhancement of Speech Processing Course. Funded by the Ministry of Education, Youth and Sports of the Czech Republic. Grandholder: Hicham Atassi

2008-2010 C08057 - Analysis and Enhancement of Speech and Image Signals form Noise for Cross-Modal Analysis of Verbal and Non-verbal Communication. Grant holder: Zdenek Smekal.

2007-2009 JEP 2006 - New MSc Curriculum in TeleInformatics. Grant holder: Zdenek Smekal. Funded by EU

2007-2009 FT-TA2/072 - Research and application of time-frequency analysis in logopaedia, foreign language learning and learning speech of deaf people. Funded by the ministry of industry and trade of the Czech Republic. Grantholder: Kamil Vrba,

Paper review

International Journal of Artificial Intelligence Tools (3 papers)

Elsevier Journal of Computer Speech and Language (1 paper)

Lecture Notes in Computer Science (4 papers)

International Conference on Telecommunications and Signal Processing TSP (7 papers)

Elektrorevue- internet journal (2 papers)

Springer Journal of Cognitive Computation (2 papers)

Registered products

- ATASSI, H.: ARES01; Software application for emotion recognition from speech
- ATASSI, H.; PŘINOSIL, J.: ARES02; System for speaker's emotional state recognition
- ATASSI, H.; KUREČKA, R.; SYSEL, P.: DPVv1. 7; Detector of fault pronunciation
- KUREČKA, R.; SYSEL, P..; ATASSI, H.: LDNv1. 5; Labeled speech database for the detection of false pronunciation
- ATASSI, H.; PŘINOSIL, J.; MÍČA, I.; VRBA, K.; SMÉKAL, Z.: Emoce Retia 2011; Software application for speaker's emotion recognition for Czech, Slovak, Polish, Russian, Italian and French
- MÍČA, I.; PŘINOSIL, J.; ATASSI, H.; VRBA, K.; SMÉKAL, Z.: VAD Retia 2011; Voice activity detection module

Patents

ATASSI, H.; PŘINOSIL, J.; MÍČA, I.; VRBA, K.; SMÉKAL, Z.; Retia a.s., Pardubice - Zelené předmestí, CZ , Vysoké ucení technické v Brne, Brno, *Multilingual speech analyzer for the recognition of emotion, age and gender*

Publications

-ATASSI, H.; SMEKAL, Z.. Automatic Identification of Successful Phone Calls in Call Centers Based on Dialogue Analysis. ". *In Proceedings of 5rd IEEE International Conference on Cognitive Infocommunications* (submitted)

-FAROOQ, K.; KARASK, J.; ATASSI, H. A Novel Cardiovascular Decision Support Framework for Effective Clinical Risk Assessment. *In 2014 IEEE Symposium on Computational Intelligence in healthcare and e-health (IEEE CICARE 2014).* 2014 (Accepted)

-VYAS, G.; DUTTA, M.; ATASSI, H.; BURGET, R.;. Detection of chorus from an audio clip using dynamic time warping algorithm. *In Engineering and Computational Sciences (RAECS), 2014 Recent Advances in.* 2014. s. 1-6. ISBN: 978-1-4799-2290- 1.

-DUTTA, M.; SINGH, A.; BURGET, R.; ATASSI, H.; CHOUNDHARY, A.; SONI, K. Generation of biometric based unique digital watermark from iris image. In *36th International Conference on Telecommunications and Signal processing.* 2013.s. 808-812. ISBN: 978-1-4799-0402- 0.

-KOPŘIVA, T., ATASSI, H., HUSSAIN, A. Classification of Transmission Channels by Speech Signal Processing. *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, 2–4 July 2013.

-FAROOQ, K., HUSSAIN, A., ATASSI, H., LESLIE, S., ECKEL, C., MACRAE, C., & SLACK, W. (2013). A Novel Clinical Expert System for Chest Pain Risk Assessment. In *Advances in Brain Inspired Cognitive Systems* (pp. 296-307). Springer Berlin Heidelberg.

-ATASSI, H.; MÍČA, I. The influence of speakers emotional state on the gender recognition process. *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz),* 2012, vol. 2012, no. 12, p. 1-5. ISSN: 1213- 1539.

-ATASSI, H.; SMÉKAL, Z.; ESPOSITO, A. Emotion Recognition from Spontaneous Slavic Speech. In *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012).* 2012. p. 389-394. ISBN: 978-1-4673-5185- 0.

-ATASSI, H.; HUSSAIN, A.; SMÉKAL, Z. Find My Emotion in the Space: A Novel Approach to Vocal Emotion Recognition. In*6th International Conference on Teleinformatics.* 2011. p. 230-235. ISBN: 978-80-214-4231- 3.

-KUBÁNKOVÁ, A.; ATASSI, H.; ABILOV, A. Selection of Optimal Features for Digital Modulation Recognition. In*Proceedings of the 10th WSEAS International Conference on System Science and Simulation in Engineering (ICOSSSE 11).* Penang, Malaysia: WSEAS Press, 2011. p. 229-234. ISBN: 978-1-61804-042- 8.

-KUBÁNKOVÁ, A.; ATASSI, H.; KUBÁNEK, D. Noise Robust Automatic Digital Modulation Recognition Based on Gaussian Mixture Models. In *Proceedings of the 6th International Conference on Teleinformatics - ICT 2011 (id 18951).* Brno, Czech Republic: VUT v Brne,

2011. p. 220-226. ISBN: 978-80-214-4231- 3.

-KUBÁNKOVÁ, A.; ATASSI, H.; KUBÁNEK, D. Gaussian Mixture Models- based Recognition of Digital Modulations of Noisy Signals. *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz),* 2011, vol. 2, no. 1, p. 15-22. ISSN: 1213- 1539.

-ATASSI, H.; ESPOSITO, A.; SMÉKAL, Z. Analysis of High- level Features for Vocal Emotion Recognition. In *34th International Conference on Telecommunications and Signal Processing.* 2011. p. 361-366. ISBN: 978-1-4577-1409- 2.

-MÍČA, I.; ATASSI, H.; PŘINOSIL, J.; NOVÁK, P. Voice activity detection under the highly fluctuant recording conditions of call centres. In *Advances in Communications, Computers, Systems, Circuits and Devices.* 2010. p. 334-336. ISBN: 978-960-474-250- 9.

-ALI, R.; HUSSAIN, A.; ATASSI, H. Intelligent Signal Image Processing Techniques for Aquaculture Application. In *SICSA PhD Conference 2010.* Edinburgh, UK: 2010. p. 59-62. ISBN: 0-02-919235- 8.

-SMÉKAL, Z.; ATASSI, H.; STEJSKAL, V.; MEKYSKA, J. Hidden Markov Model Toolkit (HTK). *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz),* 2009, vol. 2009, no. 11, p. 11- 1 (11-42 p.)ISSN: 1213- 1539.

-BENEŠ, R.; ATASSI, H.; ŘÍHA, K. Real- Time Digital Image Segmentation and Object Classification. In *32nd International Conference Proceeding on Telecommunications and Signal Processing - TSP' 2009.* Budapest, Hungary: Asszisztencia Szervezo Kft., 2009. p. 70-74. ISBN: 978-963-06-7716- 5.

-ATASSI, H.; RIVIELLO, M.; SMÉKAL, Z.; HUSSAIN, A.; ESPOSITO, A. Emotional Vocal Expressions Recognition using the COST 2102 Italian Database of Emotional Speech. *Lecture Notes in Computer Science,* 2009, vol. 2009, no. 5967, p. 1-14. ISSN: 0302- 9743.

-KOUŘIL, J.; ATASSI, H. Objective Speech Quality Evaluation. A primarily Experiments on a Various Age and Gender Speakers Corpus. In *Proceedings of The 8th WSEAS International Conference on CIRCUITS, SYSTEMS, ELECTRONICS, CONTROL & SIGNAL PROCESSING.* Puerto De La Cruz, Spain: WSEAS Press, 2009. p. 333-336. ISBN: 978-960-474-139- 7.

-ATASSI, H.; ESPOSITO, A. A Speaker Independent Approach to the Classification of Emotional Vocal Expressions. In *Proceedings of Twentieth International Conference on Tools With Artificial Intelligence, ICTAI 2008.* Dayton, Ohio, USA: IEEE Computer Society, 2008. p. 147-152. ISBN: 978-0-7695-3440- 4.

-ATASSI, H. Fundamental frequency detection methods. *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz),* 2008, vol. 2008, no. 4, p. 1-17. ISSN: 1213- 1539.

-ATASSI, H.; SMÉKAL, Z. Real- Time Model for Automatic Vocal Emotion Recognition. In *Proceedings of 31th International Conference on Telecommunications and Signal Processing - TSP 2008.* 2008. p. 90-95. ISBN: 978-963-06-5487- 6.

Number of citations (without self-citation): 29

H-index according to Google scholar: 4

Research interests

Speech processing, image processing, pattern recognition, artificial intelligence, human-machine interaction, affective computing, physiological signal processing, healthcare expert systems, robotics

Technical skills

Working knowledge	Matlab/Simulink, Octave, C/C++, Assembler, Java, Python, RapidMiner, Weka, Praat, IBM SPSS, AutoIt, DSP programming (Freescale and Motorola), Arduino, computer networks and Windows Servers, Photoshop
Basic knowledge	C#, PHP, Opnet Modeler, Maple, Mathematica, Eagle, AutoCAD

Hobbies

Reading, football, tennis, bowling, photography, graphic design, RC planes and flight simulators