VYSOKÁ ŠKOLA BÁŇSKÁ–TECHNICKÁ UNIVERZITA OSTRAVA
VŠB–TECHNICAL UNIVERSITY OF OSTRAVA

FAKULTA ELEKTROTECHNIKY A INFORMATIKY
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

KATEDRA TELEKOMUNIKAČNÍ TECHNIKY
DEPARTMENT OF TELECOMMUNICATIONS

Klasifikace emocí v lidské řeči

Classification of Emotions in Human Speech

DIZERTAČNÍ PRÁCE
DISSERTATION THESIS

AUTOR PRÁCE
AUTHOR
Ing. Pavol Partila

VEDOUCÍ PRÁCE
SUPERVISOR
doc. Ing. Miroslav Vozňák PhD.

OSTRAVA, 2016

DECLARATION

I declare that I have written my doctoral thesis on the theme of “Classification of emotions in human speech ” independently, under the guidance of the doctoral thesis supervisor and using the 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.

As the author of the doctoral thesis I furthermore declare that, as regards the creation of this doctoral thesis, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal and/or ownership rights and I am fully aware of the consequences in the case of breaking Regulation § 11 and the following of the Copyright Act No 121/2000 Sb., and of the rights related to intellectual property right and changes in some Acts (Intellectual Property Act) and formulated in later regulations, inclusive of the possible consequences resulting from the provisions of Criminal Act No 40/2009 Sb., Section 2, Head VI, Part 4.

. . . .

(author's signature)

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my supervisor Assoc. Prof., M.Sc. Miroslav Voznak, Ph.D. for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I thank my fellow colleagues for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last five years. In particular, I am grateful to M.Sc. Jaromir Tovarek for his untiring cooperation and help with research directly associated to this work.

Last but not the least, I would like to thank my family: my parents and to my brother and sister for supporting me spiritually throughout writing this thesis and my life in general.

. . . .

(author's signature)

ABSTRAKT

Dizertačná práca sa zaoberá problematikou rozpoznania emočného stavu z reči človeka. Práca popisuje súčasný stav problematiky Speech Emotion Recognition, zaoberá sa metódami na extrakciu rečových príznakov, klasifikačnými metódami a je venovaná návrhu nového systému pre klasifikáciu emočného stavu z reči. Tento systém je namodelovaný na novovytvorenej emočnej databáze emoDBova a databáze pre detekciu stresu 112DB a implementovaný do infraštruktúry zabezpečeného komunikačného systému. Nové databázy sú vytvorené z spontánnej reči v českom jazyku. Systém pre rozpoznávanie emočného stavu je navrhnutý na základe posledných poznatkov a za účelom dosiahnutia vyššej presnosti ako prezentujú doterajšie návrhy. Celý systém je implementovaný do spomínanej infraštruktúry za účelom rozpoznávania emočného stavu účastníkov telefónneho rozhovoru. Spomínané novovytvorené databázy, unikátny systém pre rozpoznanie emočného stavu a jeho reálne nasadenie v komunikačnej infraštruktúre sú hlavnými prínosmi tejto práce.

KLÍČOVÁ SLOVA

Reč, emočný stav, klasifikácia, rozpoznanie, česká databáza.

ABSTRACT

Dissertation thesis deals with recognition of the emotional state from human speech. The dissertation describes the current state of the Speech Emotion Recognition topic, deals with methods for speech features extraction, classification methods and is devoted to the design of a new system for speech emotion recognition. This system is modeled on the newly created emotional database emoDBova and the new database for stress detection 112DB. Designed speech emotion recognition system is implemented in secure communication infrastructure. The new databases are composed of spontaneous speech in the Czech language. The system for speech emotion recognition is designed on the basis of the last knowledge and to achieve higher accuracy than relevant proposals. The system is implemented to infrastructure, and its role is speech emotion recognition of phone call participants. Above mentioned newly created databases, a unique system for speech emotion recognition and its actual implementation in communications infrastructure are also major contributions of this work.

KEYWORDS

Speech, emotional state, classification, recognition, Czech database.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations

ACF	Autocorrelation Function
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
AUC	Area Under Curve
CCA	Canonical Component Analysis
CPP	Cepstral Peak Prominance
DCT	Discrete Cosine Transform
FFBP-NN	Feed-Forward Back Propagation Neural Network
GMM	Gaussian Mixture Models
HLDs	High-Level Descriptors
HMI	Human-Machine Interaction
HMM	Hidden Markov Model
IDFT	Inverse Discrete Fourier Transform
IRS	Integrated Rescue System
k-NN	k-Nearest Neighbors
LLDs	Low-Level Descriptors
LSP	Line Spectral Pairs
MCS	Multi-Classifier Systems
MFCC	Mel-frequency Cepstral Coefficients
MSE	Mean Squared Error
PCA	Principal Component Analysis
ROC	Reciever Operating Characteristic
SCG	Scaled Conjugate Gradient
SER	Speech Emotion Recognition
SHS	Sub-Harmonic Sampling
SVM	Support Vector Machine

VQF Voice Quality Features

ZCR Zero Crossing Rate

Symbols

α constant of filter steepness

$\bigtriangledown E$ Gradient of MSE

γ Impact constant of previous sample

μ_s Mean value of speech signal

$\triangle c_m$ Delta coefficients of cepstrum

$\triangle^2 c_m$ Delta-delta coefficients of cepstrum

A_{CM} Accuracy - from Confusion Matrix

$c(n)$ Cepstral coefficients of signal $x(n)$

$D(i)$ i^{th} classifier

$d\left(X, Y\right)$ Various distances between the vector x_i and y_i

E Temporal energy

$E(w)$ Mean squared error of weights

F_0 Fundamental frequency

f_{mel} Mel-frequency scale

F_{score} Harmonic mean of precision and recall

F_s Sampling frequency

FN False Negative

FNR False Negative Rate

FP False Positive

FPR False Positive Rate

$H(m)$ Harmonicity of m-th segment

$H(z)$ Transfer function of FIR filter

k Lag

l_o Overlapping length

l_s The length of segment

N	The length of signal or segment
N_s	Number of segments
PPV	Precision, Predicted Positive Rate
$R(m)$	Autocorrelation function
R_0	First coeff. of ACF - energy
R_{max}	Maximum of ACF
$s(n-m)$	Shifted signal $s(n)$
TN	True Negative
TNR	Specificity, True Negative Rate
TP	True Positive
TPR	Sensitivity, True Positive Rate, Recall
$w(n)$	Hamming window
$X(i)$	Spectrum of speech signal
$b_j^{(1)}$	Bias parameters associated with the hidden units
$w_{ij}^{(1)}$	Elements of first-layer weight matrix

1 INTRODUCTION

Recent decades have brought results which show that electrical engineering fields could not move forward without development in Signal Processing. One indicator of technological progress has been enormous efforts to simplify human-machine communication. The way of exchange and obtain information becoming user-friendly, that means using resources such as speech or visual expression. This is the cause of numerous scientific communities dealing with speech and image processing. Speech processing could be divided into three application areas:

- speech recognition (Speech-To-Text)
- speech reconstruction (Text-To-Speech)
- speaker recognition (recognition of identity, gender, age, emotions and so on)

From the title of thesis, it is evident that this work is dedicated to the speech emotion recognition and the motivation for the research is based on the concern to improve existing systems used for speech processing.

Regarding analysis, machine part must be able to extract information from human speech and visual expression. Until recently, scientific teams pay attention to the content extraction. That is, to determine what a person says. Their methods bring relatively high accuracy of applications such as Speech-To-Text (depending on the language) and others. The next step in the analysis is the extraction of secondary information. It means the issue is how the information is expressed. The aim is to find out in what situation a speaker is. For human speech, it is possible to find out information such as age, gender, identity and even emotional state. This type of information provides a new dimension in human-machine communication. Concerning synthesis, there are applications such as Text-To-Speech and Facial Expressions in robotics. Especially in this case, information about the emotional state of the speaker enables the machine to change the mood of communication.

Usability of information about the emotional state is wide. According to [1], "if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even to have and express emotions." From this perspective, an emotional condition is utilized to estimate the mood and adapt the computer's response to the man. Further use is the detection of stress [2]. Men in action such as an army, police, and fire components are exposed to high psychological pressure. Dispatching can use the information extracted from communication channel about the agent under stress and changes actions tactics and procedure. With information about customer's emotion state, an agent in call management or marketing centers will assess the situation and respond appropriately on customer reaction [3]. On the other hand, the success of agents and products also can be rated by customer satisfaction. We can assume that speech processing can detect irony and sarcasm like text-dependent systems used in social media [4], which gives the opposite of the content. The phrase "It is really fun." changes the meaning with ironic expression. Speech Emotion Recognition (SER) can be also used in lie detection. An untrained person can simulate and control the emotions tough. Accurate emotional state detection and its variations constitute a change in human behavior and can prove a lie from speech.

The motive for SER was established, on the other hand, the task of speech emotion recognition is a big challenge for several reasons. First, it is not clearly stated which speech features are useful and significant in distinguishing between emotional states. Sound diversity caused by the existence of differences in sentence structures, speakers, conversational style and rate of speech brings more difficulties because these characteristics influence the basic parameters such as fundamental frequency (F_0) and energy [5]. Also, sometimes conversational speech contains phrases in which there is more than one emotion. Another challenge is how a particular emotion expressed by individual speakers. Different cultures and nations express emotions with a different intensity. A greater amount of work is focused on the monolingual database where those differences do not exhibit so markedly. On the other hand, some works deal with SER for multilingual speech, such as [6]. Another problem is the long-term effects of emotions as in the case of sadness, and wherein the stimulation time can be as days to weeks. In a short period (minutes) can the depressed person fall to other emotional states. Unfortunately, it is not set out how the system recognizes the emotional state. It depends on the definition of emotional states that we want to detect. This is a crucial part of the development of a system for detecting the emotional status of the speech. Given the emotion expression and its effect is not generally defined, it is necessary to determine this before the development of the SER system and given its application field.

Theorists define a wide range of emotions. The issue of SER limits the work of scientific teams. Most research has been devoted to classification only the emotions such as neutral, anger, fear, disgust, joy, sadness and surprise. These emotions are most significant in human speech [7].

This thesis is organized in following way:

- Chapter 2: State of the art - describes the current status of the issue. Findings are strictly linked to the related research.
- Chapter 3: Goals of dissertation thesis - defines specific objectives for whole research.
- Chapter 4: Applied approaches to Feature Extraction - describes methods used for feature extraction. The chapter also includes a listing of used features.
- Chapter 5: Classification methods applied in system design - chapter describes implemented classification methods and the evaluation of their accuracy.
- Chapter 6: First draft of classification system and evaluation on reference database - describes BerlinDB and experiment investigating the accuracy of classification methods. The second experiment contains the first draft of a classification approach for recognizing emotional states.

- Chapter 7: Experimental evaluation of classifiers on new created databases - creating a new emoDBova database is described in this chapter. The classification accuracy of used methods is examined and compared with BerlinDB results. The end of the chapter is devoted to additional databases.
- Chapter 8: Proposal of Multi-Classifier SER System and verification of new approach - describes the design approach for classifying the emotional state from the Czech speech.
- Chapter 9: Experimental implementation of of proposal system in secured communication infrastructure - describes the application of the proposed classification approach to the real communication system.

2 STATE OF THE ART

In general, recognizing emotional states from human speech requires a similar approach to any pattern recognition issue. The vast majority of previous approaches generalized procedure for estimation of emotional speech. The block diagram in Fig. 2.1 shows the main steps of the procedure.

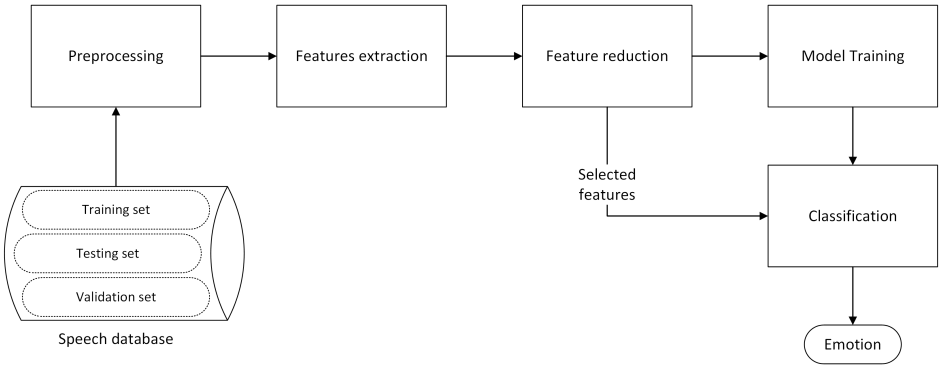


Fig. 2.1: Block diagram of the SER system. The process is divided into training, testing, and validation phase. Selected features are used to test and validate the system.

This section presents a summary overview of three major components SER:

- Theory of human emotions.
- Design criteria of speech databases.
- Influence of feature selection on classification accuracy.
- Classification methods used for emotion recognition.

2.1 Human emotion

Emotion is a mentally and socially complex reaction of an organism to a significant object or event influenced by hormones. These responses are manifested with physiological changes in the heart rate, respiration frequency, motor symptoms and changes in promptness, concentration and reaction time. Stress is a part of emotion which is the running condition of any living organism exposed to unusual situations (mental or physical) with subsequent defensive reactions which aim to maintain homeostasis and prevent the damage or death of the organism [9]. There are many areas in which the information about the emotional state is needed. Nowadays, technological development puts more emphasis on the increased accuracy and simplicity of communication between man and computer. Modern applications use the speech for input-output interface increasingly. In this type of interaction, two problems can occur, caused by the absence of information about the emotional state. The first one is an incorrect recognition of a word or a command from a person who is under stress. The machine recognizes human speech differently than a man with his hearing. The accuracy is affected by changes in the voice signal due to stress on the vocal tract. The second problem is that we feel the absence of the emotional state in the machine speech of the loudspeaker. Classic applications such as Text-To- Speech combine parts of speech sounds that are truly correct, but ultimately this signal is without any emotion. Such speech acts on the man and is synthetically unreliable. There are several physiological criteria such as heart rate, breathing changes, and sweating, which enable determining the emotional state of a man. Some speech signal parameters are used in speech processing. An imperfect human ear responds to parameters such as intensity, intonation, and speech rate. The fundamental frequency of speech, zero crossings rate, energy, and cepstral coefficients are parameters which are used in digital speech processing.

Features delimitation depended on affective states are shown in the table below.

Tab. 2.1: Feature delimitation of different affective states [10]. 0 low; + medium; ++ high; +++ very high; - range.

type of affective state	**intensity**	**duration**	**synchronization**	**event focus**	**appraisal elicitation**	**rapidity of change**	**behavioral impact**
Emotion	++ - +++	+	+++	+++	+++	+++	+++
Mood	+ - ++	++	+	+	+	++	+
Interpersonal stances	+ - ++	+ - ++	+	++	+	+++	++
Attitudes	0 - ++	++ - +++	0	0	+	0 - +	+
Personality traits	0 - +	+++	0	0	0	0	+

The emotional theory defines what emotions are, their number and their distinctness. New emotional theories are strongly influenced by Darwin and James [11]. Given the difficulty to extract the emotional state of human speech, the vast majority of research classifies few best recognizable emotional states. Therefore, systems for detecting emotional state are based on the basic emotional models with few stronger emotions.

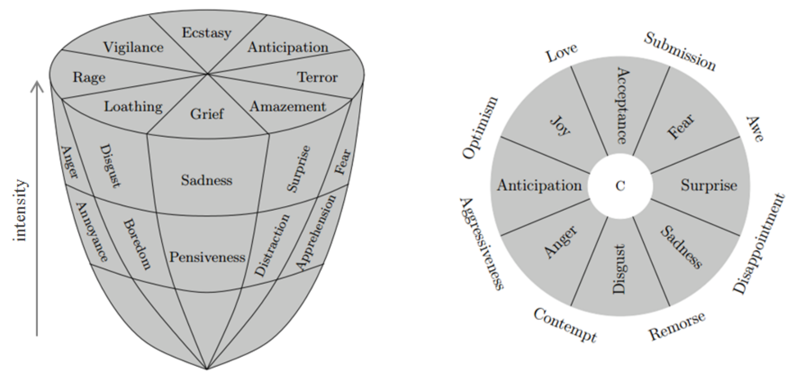


Fig. 2.2: Plutchik's emotion solid and Plutchik's emotion wheel. [12]

Figure 2.2 shows Plutchik's emotion solid and wheel. The vertical direction of solid represents the intensity of emotional state. It should be noted that the lower the intensity, the harder it is to separate emotion and recognize it. Therefore, some studies about emotion recognition discussed the issue about the recognition of stress i.e. detect the emotional state different from neutral. In this case, the neutral state is the reference against others emotions (stress) which are separated vertically to active or passive and horizontally to negative or positive against emotional reflection. This model is shown in figure bellow.

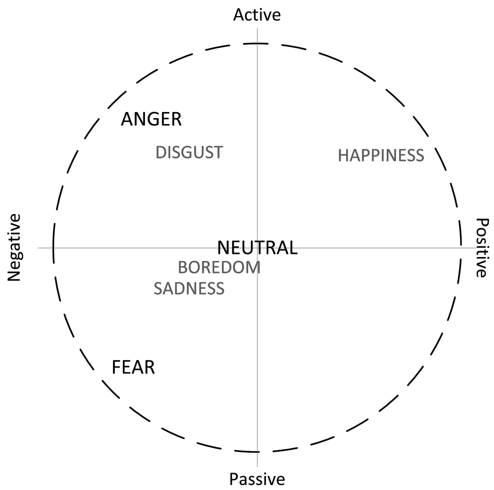


Fig. 2.3: Circle of emotions with neutral centroid.

2.2 Emotional speech databases

Number emotional speech databases are quite large. They differ on a number of factors which must be taken into account especially for comparing SER systems with each other. Listing of 64 emotional speech database is shown in Tab. A.1 located in Appendix A [13]. Additional information such as language, number, and nature of the subject, additional physiological signals related to emotional states, the purpose of the database (analysis or synthesis), contained emotional states, and the kind of emotions (natural, simulated and estimated). At first glance, it can be seen that SER systems are limited to the extraction of several emotional states. Most corpora consist of a maximum six emotional states. Figure 2.4 shows 3-dimensional emotion space with 6 emotional states [56].

These emotions are induced by the real life stimulus except for a few which come from soldiers or passengers. For example, many words with an emotional undertone, initially found in the semantic Atlas of Emotional Concepts, are referred in [15]. The pallet theory was designed to define all emotions as a mixture of some basic emotional states like the color spectrum [14]. This theory was discarded by Eckman [16] and the concept of basic emotions was extended without the assumption of mixing primary emotions. Eckman set 17 basic emotional states. Databases of this type rarely contain emotions out of this list, and we call them the higher-level emotions [17].

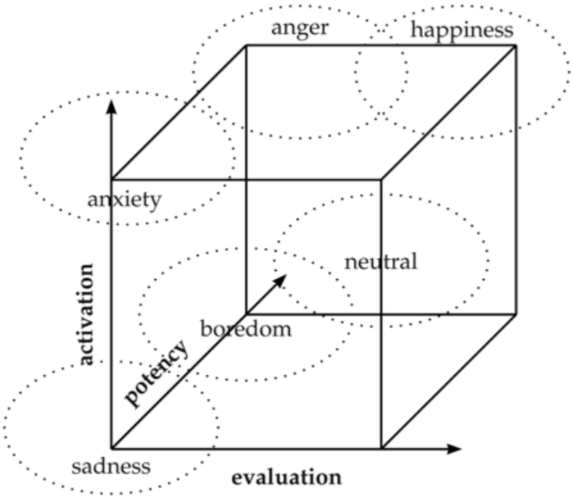


Fig. 2.4: 3-dimensional emotion space with six emotional states [56].

The database can consist of three emotion types. Natural speech contains emotions that are induced by real life stimuli. Simulated (acted) speech contains a purposely induced emotional states. In many cases, speech comes from professional actors who can reliably simulate emotional state. The third type of emotions is elicited. Psychology shows that we can influence a person to change the emotional state. This is useful for creating a database with the lack of actors [18]. Acted speech has many advantages. The speech signal are obtained from a recording studio (no noise), the composition of the subjects is adjustable, the content of emotional states is evenly distributed, also simulated emotions is often excessive and therefore easily distinguishable. On the other hand, is not correct to train the system for acted speech and use it for speech emotion recognition in the real environment.

Along with the speech signals, some databases contains other physiological information. The intensity of breathing, heartbeat, systolic and diastolic blood pressure, even sweat on skin indication constitute additional data are added to speech corpus. These physiological data are used mainly to evaluate the degree of stress stimuli [19], [20], [Par14c]. Research presented in [21] points out that mentioned additional data are much more related to the excitement of subject speech and do not notice such a value for determining the particular emotional state. For example, happiness can cause increased heart activity as well as anger. This finding also applies to the EEG signal [8]. There is a correlation between the change in the emotional state and the EEG signal, nevertheless is not possible to determine the particular emotional state.

Recorded data collection is useful to compare the results of other experiments. The new database can be created from the media stream (radio, television) [15]. In this case, it may be a problem with copyright. Another type of speech is an interview with specialists such as psychologists or scientific specialists on phonetics [15]. Real-life situations such as interviewing employees for promotion is used in [19]. Parents educate children and alert them from dangerous situation made a speech source for [22]. The reactions of patients to explain the doctor diagnosis is another source of speech for this use [23]. In [24], speech is recorded with Human-Machine Interaction (HMI) where the Automatic Speech Recognition (ASR) is used during a phone call.

Every emotional speech database has some limitations. The system does not achieve sufficient performance with limited training data. These limits are (in brief):

- Many databases contain samples that are not sufficiently stimulated emotionally. In this case, the subjective recognition of emotion in this kind of database has a lower accuracy (65 percent of personal recognition rate in [25]).
- In some databases such as [26] recordings have very low sound quality
- Some database does not contain phonetic transcription [22], which is undesirable for the content-dependent systems.

2.3 Feature extraction and selection methods

The SER is in most cases performed by speech processing without linguistic information. Present knowledge allows speech processing to extract valuable information from the pure acoustic signal. However, there are cases where SER is supported by the ASR. At first glance, ASR enhances the classification with linguistic information where the mood of conversation can be estimated from content. This fact is widely misleading. The vast majority of the ASR is developed and worked on speech databases which do not contain spontaneous speech. Speech recognition is not used due to the inability to recognize the content (linguistic information) of the emotionally tuned speech. After the speech processing and feature extraction is a standard step of selecting symptoms. After the speech processing and feature extraction is advisable to select only significant features to reduce feature vector to ”*golden set*” and accelerate classification. The highlighted part of Fig. 2.5 shows the basic procedural pipeline which marks the SER and ASR separately.

2.3.1 Non-linguistic features

The group of features (feature vector) is divided into segmental and suprasegmental according to its temporal representation of the original signal. Segmental symptoms (short-term acoustic features) are extracted from short frames, mostly 25–50ms. On the other hand, suprasegmental features are calculated from a much wider frame such as the whole utterance as seen in Fig. 2.6 [28]. Comparison of suprasegmental and segmental features is described in the Schuller and Rigoll paper [29].

Speech features are also divided into two other classes. First, Low-Level Descriptors (LLDs) contains prosodic parameters (suprasegmental) and spectral parameters and their derivatives (segmental features). Second class, functionals or High-Level Descriptors

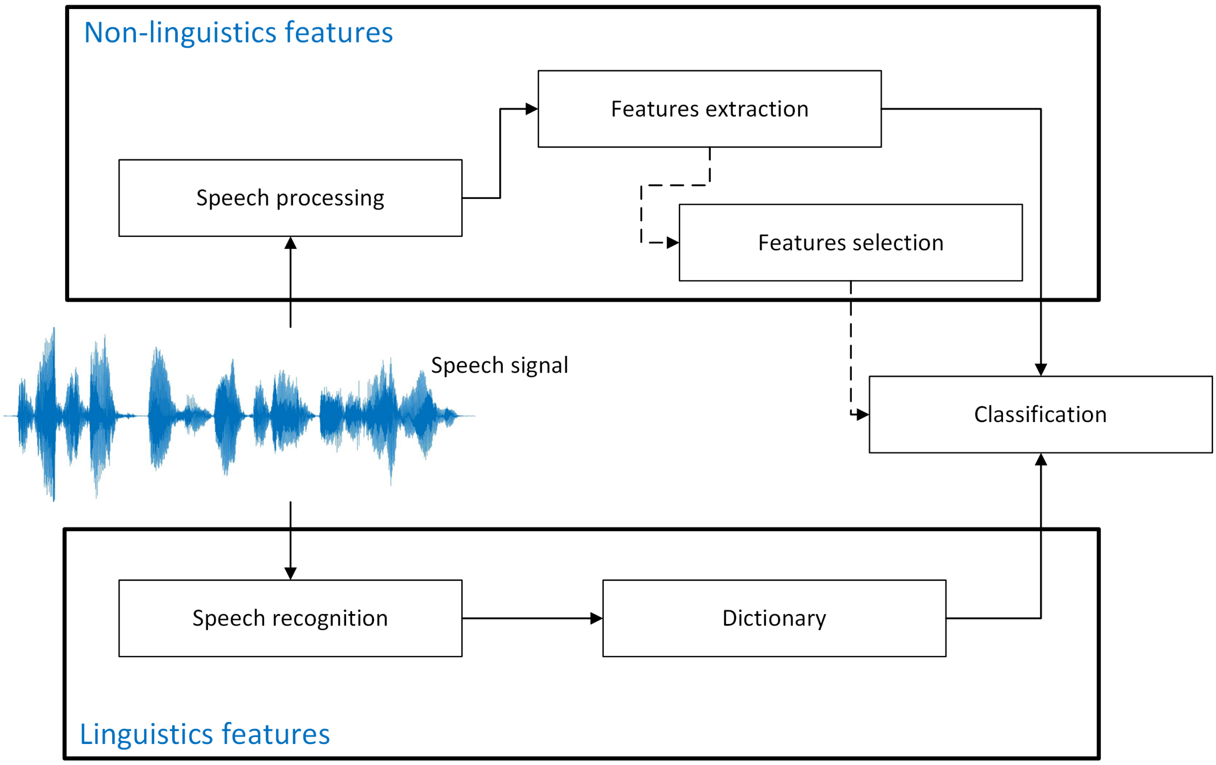


Fig. 2.5: SER and ASR processing pipeline.

(HLDs) represents statistical derivatives of LLDs and belongs to suprasegmental features. Table 2.2 contain descriptions of speech features and Tab. 2.3 shows LLDs and functionals.

Tab. 2.2: Speech features and description. [39]

Features	**Description**
Mel-frequency cepstral coefficients (MFCCs), Linear prediction cepstral coefficients (LPCCs)	Derive from cepstrum, which is inverse spectral transform of the logarithm of the spectrum
Formants (spectral maxima or spectral peaks of the sound spectrum of the voice), log-filter-power-coefficients (LFPCs)	Derive from Spectrum
Noise-to-harmonic ration, jitter, shimmer, amplitude quotient, spectral tilt, spectral balance	Are measurements of Signal (voice) quality
Energy, short energy	Are measurements of intensity
Fundamental frequency (pitch)	Are measurements of frequency
Temporal features (duration, time stamps)	Are measurements of time

Previous studies have dealt with prosodic symptoms such as F_0, duration and intensity. Relatively small feature vectors were extracted (10-100 features) [30], [31], [32], [33]. More current research since 2007 until now points to the trend of speech features as HNR, jitter, shimmer, and all segmental features from Tab. 2.3 [34], [35], [36], [37].

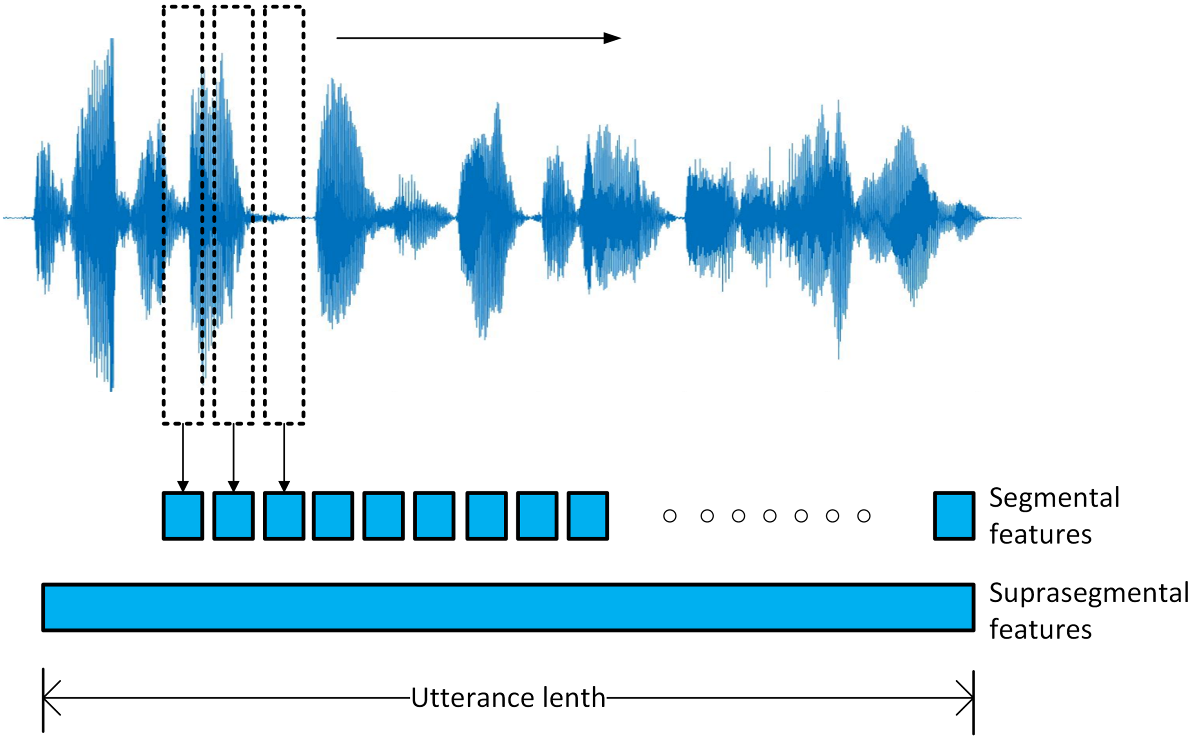


Fig. 2.6: Segmental (short-term) and suprasegmental (long-term) features extracted from speech signal.

Tab. 2.3: Segmental and suprasegmental speech features categorized to LLDs and functionals (HLDs). [39]

Low-level descriotors (LLDs)	**functionals (High-Level Descriptors)**
Suprasegmental features	
Fundamental frequency (Pitch), energy, intensity, harmonic-to-noise ration (HNR), shimmer, jitter, speech rate, normalized amplitude quotient, spectral tilt, spectral balance	Extreme values (maximum, minimum), means (arithmetic, quadratic, geomteric), moments (standard deviation, variance, kurtosis, skewness), percentiles and percentile ranges, quartiles, centroids, offset, slope, mean squared error, sample values, time/durations
Segmental features	
Mel frequency cepstral coefficients (MFCCs), formant amplitude, formant bandwith, formant frequency, log-filter power coefficients (LFPCs), linear prediction cepstral coefficients (LPCCs), line spectral pairs, short (Frame) energy, frame intensity	

2.3.2 Feature selection

As mentioned, the set of features is enormous (several tens), with all the coefficients and their derivatives may reach hundreds. In the field of pattern recognition and classification is this undesired state defined as *curse of dimensionality* [38]. Acceleration of the classifier learning process requires minimizing the size of the feature vector [39]. The most frequently used methods of selecting are Principal Component Analysis (PCA) [40], [41] and Canonical Component Analysis (CCA) [42]. Correlation-based Sub Set Evaluators have also been used, where several of searching methods evaluate a subset of features.

2.3.3 Classification methods

Existing research points to many classification methods used in speech emotion recognition. Hidden Markov Model (HMM), Gaussian Mixture Models (GMM), Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), Deep-learning networks and others are most frequently used. It is clear that each type of classifier has advantages as well as disadvantages. This fact was the reason for using multiple classifiers fusion.

Each classifier reaches the various precision when processing different databases. Schuller et al. (Schu10) presented cross-corpora evaluation to increase independence between training and testing sets. There were showed results from six databases in a cross-corpora and multilingual experiment. Table 2.4 shows few important results for mentioned classifiers and standard (known) databases [39].

Tab. 2.4: Classification performance of single classifiers on well known databases.

Classifier	Performance	Reference
SVM	up to 81% in several cross-corpus experiments with varying number of classes	Schuller et al. [43]
	89% in Berlin EMO and DSPLAB databases	Yang et al. [44]
GMM	86% in Chinese LDC	Zhou et al. [40]
	81% in Berlin EMO database	Atassi and Esposito [45]
HMM	86% in Berlin EMO database	Yun and Yoo [46]
	~81% in ELSA multi-lingual emotional speech database	Nogueiras et al. [47]
ANN	~60% (speaker dependent) and 55% (gender dependent) in LDC emotional prosody speech-transcripts database	Cen et al. [48]
	83.2% (speaker dependent) and 55% (speaker independent) in Berlin EMO database	Iliou and Anagnostopoulos [49]

The hot topic in the field of pattern recognition and classification is Deep Learning. There are several studies which show promising results [50] (unknown database) or a comparison of hybrid DNN-HMM with other multiple classifiers [52], [51]. In any case, it is a challenge to determine the number of hidden layers and neurons in each layer for such networks.

3 GOALS OF DISSERTATION THESIS

The structure of the dissertation is based on the state-of-the-art research and requirements. Points mentioned bellow, represents total goals of the dissertation that describes the dissertation schedule based on the new trends in emotion recognition from speech.

- Creation of the training and testing Czech language database with various emotion state recordings.
- Design of novel classifier dealing with speech emotion recognition. The contribution will be an identification of the most significant features in speech affecting human emotions and the own classifier based on the artificial neural network.
- Verification of results and achieved contribution, compared with actual well-known systems.

Goals of the thesis have been approved by the Commission on rigorous examination, held in February 2014.

The work includes a description of the new system design for speech emotion recognition. This system should classify emotions in the Czech language. Therefore, one of the goal defines the creation of a Czech emotional database which will be used for training and testing system. The third goal defines the implementation of the system in the real environment, infrastructure for voice service (telephone calls). The implemented system will be used for analyzing of calls and classifying the emotional state of the caller. System design should meet the conditions of increased accuracy compared to the presenting proposals. It will be necessary to explore and utilize proven methods of classification. The last point defines a verification section that compares the obtained results against presented and publicized proposals.

The following chapter discusses the issue of calculation of speech parameters (feature extraction) and presents the theory regarding the speech processing.

4 APPLIED APPROACHES TO FEATURE EXTRACTION

This chapter is devoted to listing and description of the features that might be contained in the feature vector for speech emotion recognition. LLDs feature category described in Sec. 2.3 are often categorized as prosodic, spectral and voice quality features.

4.1 Pre-processing

Initial speech signal has several properties that need to be removed or modify before feature extraction. These are several operations which remove these undesirable characteristics. This obligate phase is called pre-processing [Par13].

Procedure of speech pre-processing:

1. **Sample rate conversation:** the speech signal is recorded with a different sampling frequency. Some recordings have unnecessarily high sampling frequency. Therefore it is appropriate to eliminate redundancy by resampling the original signal.
2. **DC Offset:** some audio cards and other devices add DC (Direct Current) components into the audio signal, as shown Fig. 4.1.

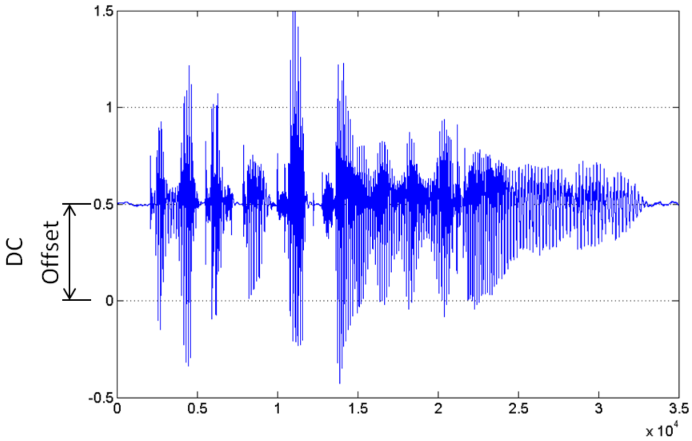


Fig. 4.1: Effect of direct current on speech signal.

The DC component in the signal negatively affects the feature extraction and may cause the disturbance. The DC component of the entire signal is computed as a mean value μ_s of all analyzed samples as shown Eq. 4.1, where N is a number of samples and $s(n)$ is a speech signal. The DC component is removed by a simple subtraction of the mean value described by Eq. 4.2. If we do not dispose of the entire signal, typically in real-time processing when a particular part of the signal is analyzed, we are not able to estimate the mean value. In this case, a real-time estimation of the mean value for each speech sample is used. The mean value for the current sample $\mu_s(n)$ from Eq. 4.3 can be determined from the mean value of the previous sample $\mu_s(n-1)$, which is linked to the actual sample by impact value γ, mostly is close to

1.

$$\mu_s = \frac{1}{N}\sum_{n=1}^{N} s(n). \tag{4.1}$$

$$s(n) = s(n) - \mu_s. \tag{4.2}$$

$$\mu_s(n) = \gamma\mu_s(n-1) + (1-\gamma)s(n). \tag{4.3}$$

3. **Normalization:** the dynamic range of the speech signal should be adjusted to the same range of -1 and 1 to avoid large differences in features (energy, and so on). Simple operation to normalize the speech signal:

$$s(n) = \frac{s(n)}{\max s(n)}. \tag{4.4}$$

4. **Pre-emphasis:** This step should be applied given the significant variations in energy in the spectral range Most of the speech signal energy is located in the first 300Hz of the speech spectrum. Since the same valuable information is also included in higher parts of the frequency spectrum, the so-called pre-emphasis is in most cases carried out by the FIR filter defined by the transfer function described in Eq. 4.5. Figure 4.2 shows speech signal before (left part) and after pre-emphasis (right part).

$$H(z) = 1 - \alpha z^{-1}, \quad \alpha \in [0.9, 1] \tag{4.5}$$

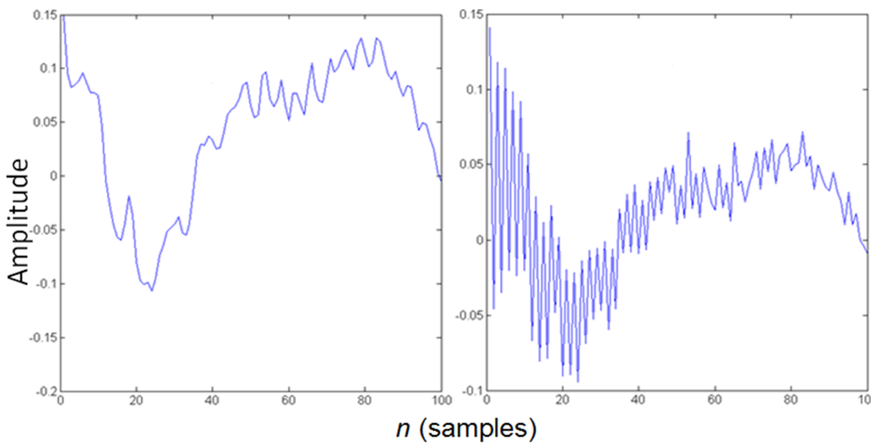


Fig. 4.2: Speech signal before and after FIR filter pre-emphasis.

5. **Segmentation:** The speech signal is non-stationary in the time domain. This attribute is undesirable for features extraction, in particular for features which define excitation from vocal tract such as F_0 and others. Therefore, it is necessary to divide speech signal into shorter segments. The segment length is usually chosen in the range of 25-50 ms with approximately half length overlapping as shown in Fig. 4.3.

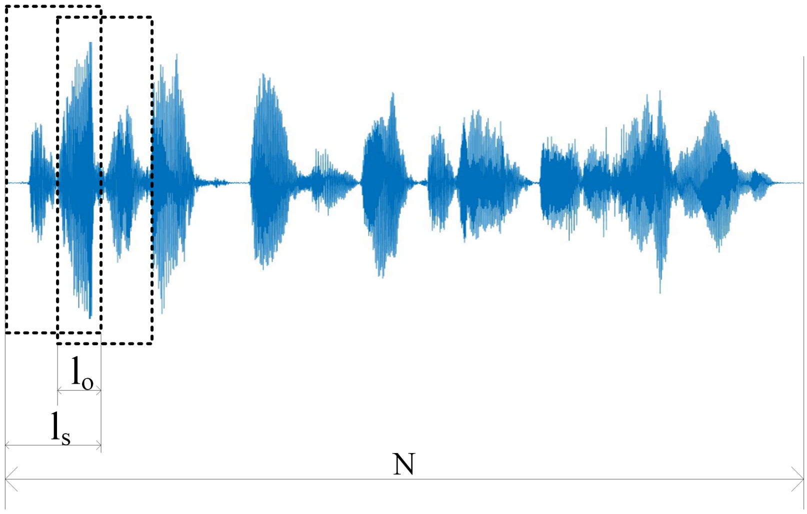


Fig. 4.3: Segmentation of speech signal with overlapping.

The number of segments N_s are calculated with formula Eq. 4.6, where l_s is a segment length, l_o is a overlap length and N is speech signal length.

$$N_s = 1 + \left\lfloor \frac{N - l_s}{l_o} \right\rfloor . \tag{4.6}$$

6. **Smoothing function:** segmentation of speech signals into frames results in a sharp transition at the edge. The sharp transition between the frames has an adverse effect, especially in frequency analysis. Multiplied window function eliminates sharp transitions on frame edges. In signal processing, many windows weighted functions can be applied. For speech recognition and to avoid the impact of sharp transitions in the spectrum, the Hamming window function is used most frequently. Equation 4.7 contains the mathematical definition of the Hamming window. The length of signal (number of samples) within one segment is represented by N, and n means a particular sample.

$$w(n) = 0.54 + 0.46 \left[\left(\frac{1}{2} N - n \right) \frac{2\pi}{N} \right] . \tag{4.7}$$

4.2 Prosodic features

Prosodic features of speech can be assessed according to several aspects. These aspects represent different representation levels of prosodic phenomena [53], [54]. Acoustic realization of prosodic phenomena can be observed and measured with different methods.

4.2.1 Energy

Signal energy is characterized by intensity of speech signal. Energy is influenced by way of recording and digitizing speech, speaker distance from the microphone, and other features. The calculation of temporal energy describes equation below.

$$E = \frac{1}{N} \sum_{n}^{N-1} (s(n))^2 . \tag{4.8}$$

4.2.2 Zero crossing rate

ZCR describes how many times speech signal change the polarity. This parameter can also carry information about variation of F_0. ZCR is calculated using sign function, as shown Eq. 4.9 below [Voz13b].

$$ZCR = \frac{1}{N} \sum_{n}^{N-1} |\mathrm{sign}\, s(n) - \mathrm{sign}\, s(n).| \tag{4.9}$$

4.2.3 Fundamental frequency

This parameter (F_0) was considered as the very important feature in speech processing. It defines excitation, the main component of speech production [55]. Age, gender, speech errors, and emotional state of a man can be determined by this feature. Vowels and consonants of speech can be precisely separated with F_0. There are several methods in signal processing which enable estimating the fundamental frequency [45]. This discipline is called "*pitch extraction*" in signal processing terminology. One of the most known pitch extraction method are based on autocorrelation function (ACF), defined by the Eq. 4.10 . The disadvantage of ACF is the inability to remove consonants. Therefore, it is appropriate to implement removing procedure (central clipping, Sub-Harmonic-Sampling (SHS) [Par12], [9] and others).

$$R(m) = \sum_{n-m}^{N-1} s(n)s(n-m), \tag{4.10}$$

and lag k is determined by Eq. 4.11 as the position of maximum (pitch) and the fundamental frequency F_0 is defined in Eq. 4.12. Figure 4.4 shows position of lag, where F_s is sampling frequency of the signal.

$$k = \operatorname*{argmax}_{m} R(m), \tag{4.11}$$

$$F_0 = \frac{F_s}{k}. \tag{4.12}$$

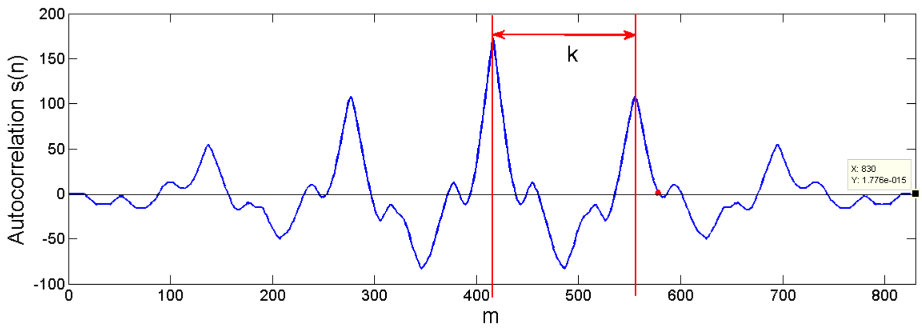


Fig. 4.4: Position of lag k from the center of ACF.

4.3 Voice quality features

Luger and Yang [56] examined the impact of VQF on SER accuracy. The results confirmed that VQF in combination with prosodic features and MFCC rapidly increase classification accuracy. Some sources report a direct link to the definition of positive and negative emotions.

4.3.1 Harmonicity

Due to the differing definitions of this parameter in some relevant literature, it is hard to determine the direct and uniform formulation. For example, the author defines harmonicity as the degree of acoustic periodicity of the speech signal [57]. On the other hand, there was an experiment in which harmonicity is defined as the dB ratio between the first maximum of ACF and signal energy [45]. 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

$$H(m) = 20 \log \frac{R_{max}}{R_0}. \tag{4.13}$$

4.3.2 Formant frequencies

Information about formant is well recognized from the spectral envelope of the analyzed speech segment. The most of the formant frequency identification procedures used implicitly or explicitly spectral envelope. Linear Predictive Coefficients (LPC) is the most common method of determining formant frequencies [58].

4.3.3 Cepstral peak prominance

Cepstral Peak Prominence (CPP) is an acoustic measure of voice quality that has been qualified as the most promising and perhaps robust acoustic measure of dysphonia severity

[59], [60]. CPP bring the information about the degree of signal harmonic organization.

4.4 Spectral features

These features can be divided into primary (basic), which use statistical view on the frequency spectrum. This group includes symptoms such as slope, skewness, etc. Second, and a more influential group contains features belonging to the homomorphic speech analysis.

4.4.1 Homomorphic speech analysis

Homomorphic analysis belongs to the group of practice non-linear signal processing, which is based on the generalized superposition principle. These procedures are suitable for analysis (separating) the signal that originated by multiplication or convolution of two or more components. The suitability of this approach stems from the model of speech signal production. A mathematical model is defined as a convolution of excitation and impulse response of the vocal tract. The aim of the analysis is determining the parameters of the system, in other words, the objective is to find and separate the individual parts of the convolution. The method of separating is also called homomorphic filtration. The following subsections will describe the features, which belong to the homomorphic analysis of speech signal.

4.4.2 Mel-frequency cepstral coefficients

For speech recognition, the most commonly used features are cepstral coefficients (MFCC especially). Cepstral coefficients are derived from an inverse discrete Fourier Transform (IDFT) of logarithm of short-term power spectrum of a speech segment as:

$$c(n) = \sum_{i=0}^{N-1} \ln\left[|X(i)|\right] e^{\frac{j2\pi ni}{N}}, \tag{4.14}$$

where $X(i)$ is the FFT-spectrum of speech $x(n)$. As the spectrum of real-valued speech is symmetric, the DFT can be replaced by Discrete Cosine Transformation (DCT). To obtain MFCC features, the spectral magnitude of FFT frequency bins are averaged within frequency bands spaced according to the Mel scale given the Eq. 4.15, which is based on a model of human auditory perception. The scale is approximately linear up to about 1000 Hz and approximates the sensitivity of the human ear.

$$f_{mel} = 2595 \log\left(1 + \frac{f}{700}\right). \tag{4.15}$$

For each vector of coefficients were derived dynamic coefficients of the delta $\triangle c_m$ and delta-delta $\triangle^2 c_m$ (acceleration coefficients), which reflect the temporal changes of coefficients vectors c_m [Tov15].

4.4.3 Line spectrum pairs

Line spectrum pairs (LSP), is a way of uniquely representing the LPC-coefficients. The motivation behind LSP transformation is greater interpolation properties and robustness to quantization. These benefits are achieved by the cost of higher complexity of the overall system. The key idea of LSP decomposition is to decompose the p-th order linear predictor $A(z)$ into a symmetrical and antisymmetrical part denoted by the polynomials $P(z)$ and $Q(z)$ respectively, as shown Fig. 4.5 below [61], [62], [63]. The LSP parameters

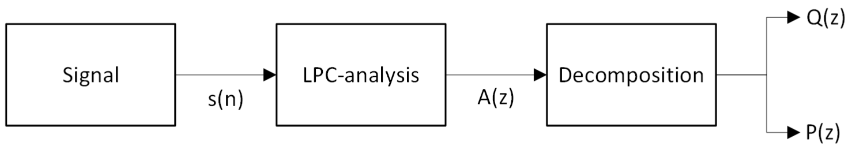


Fig. 4.5: Decomposition of the $A(z)$.

are expressed as the zeroes (or roots) of $P(z)$ and $Q(z)$. The zeroes uniquely determine $P(z)$ and $Q(z)$ and since $A(z)$ can be made up of $P(z)$ and $Q(z)$ the representation of LPC-coefficients by means of LSP-parameters is valid. The zeroes of the LSP polynomials are subject to the following properties:

1. All zeroes of $P(z)$ and $Q(z)$ are located on the unit circle.
2. The zeroes are separated and interlaced if $A(z)$ is minimum phase, i.e. $A(z)$ has all its zeroes within the unit circle.
3. All zeroes have a complex conjugate in the z-plane.

4.5 Feature vector

The ways to assemble feature vector is a lot. Mix the right combination of symptoms is a difficult task. Properties and character of the speech signal have the substantial influence on features and their significance. Drawing on previous experiments with openSMILE extractor [64], [65] was a feature vector formed from 34 low-level descriptors (LLD) with 34 corresponding delta coefficients. This feature set contains a greatly enhanced set of low-level descriptors, as well as a carefully selected list of functionals. The total number of LLD features is 68.

The list of 34 low-level descriptors:

- **pcm_loudness** The loudness as the normalised intensity raised to a power of 0.3.
- **mfcc** Mel-Frequency cepstral coefficients 0-14
- **logMelFreqBand** logarithmic power of Mel-frequency bands 0 - 7 (distributed over a range from 0 to 8 kHz)
- **lspFreq** The 8 line spectral pair frequencies computed from 8 LPC coefficients
- **F0finEnv** The envelope of the smoothed fundamental frequency contour.

- **voicingFinalUnclipped** The voicing probability of the final fundamental frequency candidate. Unclipped means that it was not set to zero when is falls below the voicing threshold.

From listed LLD features are computed 21 functionals (HLDs):

- **maxPos** The absolute position of the maximum value (in frames)
- **minPos** The absolute position of the minimum value (in frames)
- **amean** The arithmetic mean of the contour
- **linregc1** The slope (m) of a linear approximation of the contour
- **linregc2** The offset (t) of a linear approximation of the contour
- **linregerrA** The linear error computed as the difference of the linear approximation and the actual contour
- **linregerrQ** The quadratic error computed as the difference of the linear approximation and the actual contour
- **stdev** The standard deviation of the values in the contour
- **skewness** The skewness (3rd order moment)
- **kurtosis** The kurtosis (4th order moment)
- **quartile1** The first quartile (25% percentile)
- **quartile2** The second quartile (50% percentile)
- **quartile3** The third quartile (75% percentile)
- **iqr1-2** The inter-quartile range: quartile2-quartile1
- **iqr2-3** The inter-quartile range: quartile3-quartile2
- **irq1-3** The inter-quartile range: quartile3-quartile1
- **percentile1.0** The outlier-robust minimum value of the contour, represented by the 1% percentile.
- **percentile99.0** The outlier-robust maximum value of the contour, represented by the 99% percentile.
- **pctlrange0-1** The outlier robust signal range 'max-min' represented by the range of the 1and the 99% percentile.
- **upleveltime75** The percentage of time the signal is above (75% $*$ range + min)
- **upleveltime90** The percentage of time the signal is above (90% $*$ range + min)

Proces of extraction feature vector is shown in Fig. 4.6 (1428 features).

The four additional pitch related LLD (and corresponding delta coefficients) are as follows (pitch related: all are 0 for unvoiced regions, thus functionals are only applied to voiced regions of these contours):

- **F0final** The smoothed fundamental frequency contour
- **jitterLocal** The local (frame-to-frame) Jitter (pitch period length deviations)
- **jitterDDP** The differential frame-to-frame Jitter (the 'Jitter of the Jitter')
- **shimmerLocal** The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)

Besides, 19 functionals are applied to the four pitch-based LLD and their four delta coefficient contours ("pitch-based" means extraction from voiced regions of the signal only). Functionals are the set of 21 functionals mentioned above without the minimum value (the 1% percentile) and the range. Final feature vector contains from 1582 features.

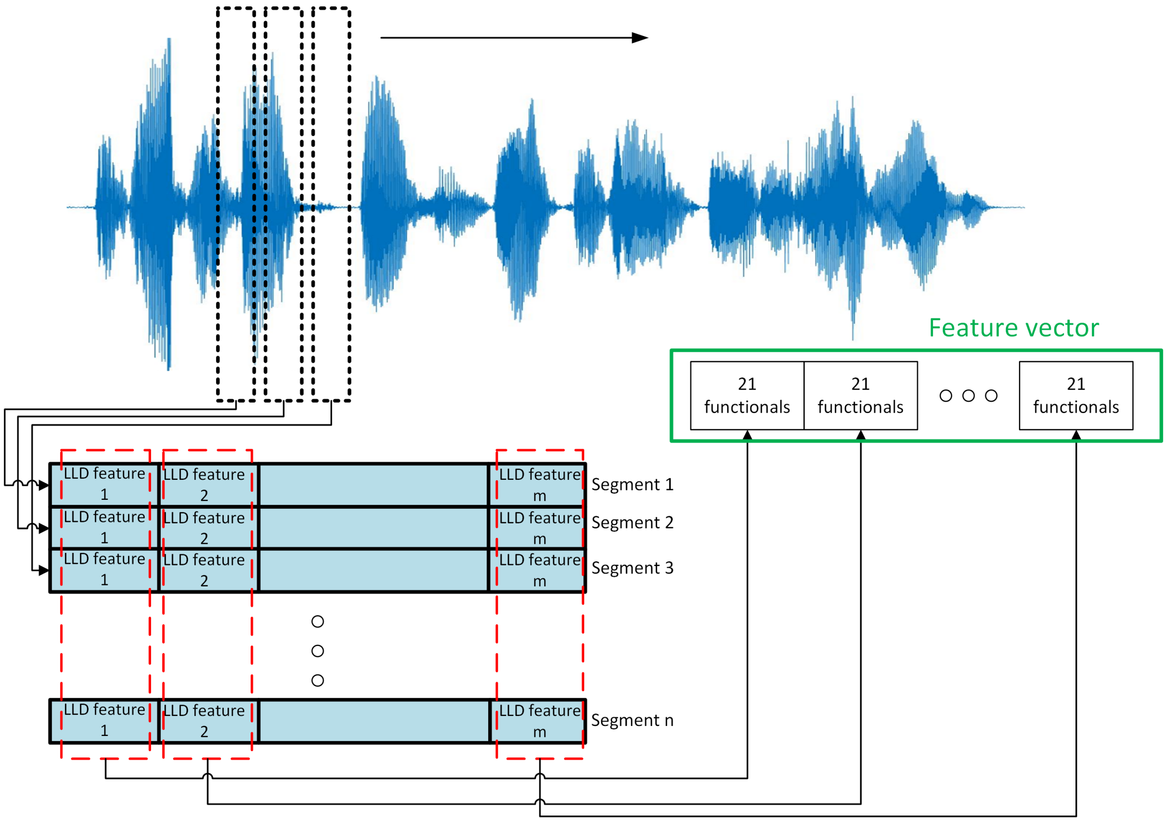


Fig. 4.6: Feature vector extraction with LLDs (m=68) and 21 functionals.

4.6 Feature selection - PCA

The role of feature selection is to reduce the dimensionality of data for classification. The size of a feature vector is directly related to the dimensionality. In general, feature selection is divided into two tasks. The first is the removal of redundant features. Some symptoms are strongly correlated and carry the same information. Therefore, it is appropriate to employ only one representative feature. Second, select most relevant features, which means take to account only most significant features due to classification task.

One of the oldest and most widely used methods of multidimensional analysis is Principal Component Analysis (PCA). The aim is to simplify definition or description linearly dependent, correlated parameters (features), and the decomposition of data into the structural matrix and noise matrix.

PCA method can be described as:

- Linear transformation of the original features to new, uncorrelated variables called principal components.
- The basic characteristic of every major component is a measure of variability or variance.
- The main components are ranked according to decreasing variance, from the highest to the lowest.
- Most of the information about the variability of the data while it is concentrated in the first component and at least the information contained in the final component.
- Most of the information about the variability of the data concentrated in the first component and at least the information included in the last component.

Definitions and deriving the principal component is available from bibliography [66], [67] and [68]

5 CLASSIFICATION METHODS APPLIED IN SYSTEM DESIGN

Individual research shows that cannot be said which classifier for emotion recognition is the best. Each classifier or the combination of classifiers achieved some results accuracy, which depends on several factors. The success of classifier is directly dependent on the data. This is derived from the fact that the accuracy varies with the data character such as the quantity, density distribution of each class (emotions) and the language also. One classifier has different results with acted database, where the density of each emotion are equitable and different with real data from call center where normal (calm) emotion state occupies 85 to 95 percent of all data. Appropriate choice of parameters has a considerable effect on the accuracy of these classifiers. The following subsections describe the used classification methods [Par14b], [Par15].

5.1 Artificial neural network

Our emotional state classification problem with a high number of parameters can be considered as a pattern-recognition problem. In this case, it can be used the two-layer feed-forward network. A two-layer feed-forward network, with sigmoid hidden and output neurons, can classify vectors arbitrarily well, given enough neurons in its hidden layer. The network is trained with scaled conjugate gradient (SCG) back-propagation. The input vectors x_i where $i = 1, \ldots, d$. The first layer of network forms M linear combinations of these inputs to give a set of intermediate activation variables $a_j^{(1)}$

$$a_j^{(1)} = \sum_{i=1}^{d} w_{ij}^{(1)} x_i + b_j^{(1)} \quad j = 1, \cdots, M, \tag{5.1}$$

with one variable $a_j^{(1)}$ associated with each hidden unit. Here $w_{ij}^{(1)}$ represents the elements of first-layer weight matrix and $b_j^{(1)}$ are the *bias* parameters associated with the hidden units. Demonstration of such a network with speech features as an input, 5 hidden layers and two output classes is shown in Figure 5.1.

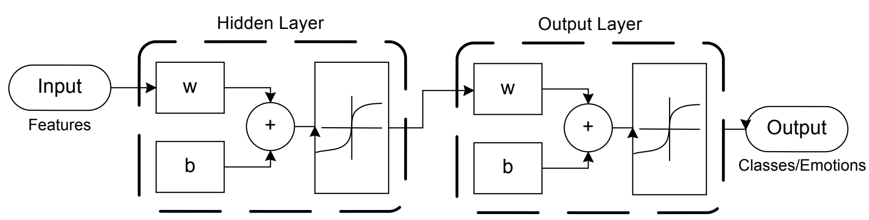


Fig. 5.1: Artificial neural network architecture with hidden layers and output classes.

SCG training implement mean squared error $E(w)$ associated with gradient $\bigtriangledown E$ and avoids the line-search per learning iteration by using Levenberg-Marquardt approach [69] to scale the step size. Weights in the network will be expressed in vector notation.

$$w = \left(..., w_{ij}^{(1)}, w_{i+1j}^{(1)}, ..., w_{N_1 j}^{(1)} \theta_j^{(l+1)}, w_{ij+1}^{(1)}, w_{i+1j+1}^{(1)}, ... \right). \tag{5.2}$$

The vector $\bigtriangledown E$ points in the direction in which $E(w)$ will decrease at the fastest possible rate. Weight update equation is shown bellow, where c is suitable constant.

$$w\,(k+1) = w\,(k) - c \bigtriangledown E. \tag{5.3}$$

The gradient descent method for optimization is very straightforward and general. Only local information, for estimation a gradient, is needed for finding the minimum of the error function. [70], [71], [Tov16].

5.2 k-nearest neighbors

The k-NN is a classification method on the principle of analogies learning. Samples from the training set are n numeric attributes, and each sample represents a point in N-dimensional space. This space of training samples is scanned by the classifier due to determine the shortest distance between training and unknown sample. Euclidean and others distances can be computed. In other words, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor. The various distances between the vector x_i and y_i.

$$d\,(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{5.4}$$

The neighbourhood distance is calculated through Euclidean metric. Given an m-by-n data matrix X, which is treated as m (1-by-n) row vectors $x_1,\ x_2, \ldots,\ x_m$.

5.3 Support vector machines

SVM offers a progressive method in the field of machine learning. The principle of classification is to find the hyperplane that divides the training data in the feature space. The optimal hyperplane is such that the training data points lie in the opposite half-space and the value of the distance between half-spaces is the largest. In other words, the goal is to maximize space among half-spaces (maximum margin). Support vectors are described by training data points that represent a decision-making role [72].

Basic tasks of the SVM is a binary classification. Ideas based on SVM (few support vectors, maximum distance and kernel transformation) were also used for the design of algorithms for other tasks. For example the role of binary classification for noisy data (soft margin), discrete classification (into several classes), regression, kernel principal component analysis (PCA), ranking, structured learning, learning from one class (one class support vector, single class data description) [73].

There is a few SVM implementation:

- Linear SVM
- Quadratic SVM
- Cubic SVM
- Gaussian SVM

5.4 Classification performance evaluation

A learned classifier has to be tested on a different test set experimentally. The experimental performance on the test data is a proxy for the performance on unseen data. It checks the classifier's generalization ability. Evaluation has to be treated as hypothesis testing in statistics.

Evaluation of the effectiveness of the classifier preceded by several principles:

- **Danger of overfitting** - Learning the training data too precisely usually leads to poor classification results on new data. Classifier has to have the ability to generalize.
- **Training vs. test data** - Finite data are available only and have to be used both for training and testing. More training data gives better generalization, and more test data gives the better estimate for the classification error probability.
- **Hold out method**
  - The data is randomly partitioned into two independent sets. Training multiset (e.g., 2/3 of data) for the statistical model construction, i.e. learning the classifier. Test set (e.g., 1/3 of data) is hold out for the accuracy estimation of the classifier.
  - Random sampling is a variation of the hold out method. Repeat the hold out k times, and the accuracy is estimated as the average of the accuracies obtained.

5.4.1 Accuracy and error rate

As mentioned, the selection of a suitable classifier is not a simple task. Therefore it is necessary to evaluate its accuracy and error. General unweighted accuracy can be calculated as:

$$A_{uw} = \frac{N_{correct}}{N_T} \cdot 100 \quad [\%], \tag{5.5}$$

where $N_{correct}$ is the number of correctly classified inputs (patterns) and N_T is the total number of inputs. In case of more classes task it is appropriate to use weighted accuracy:

$$A_w = \frac{100}{C} \sum_{c=1}^{C} N_{correct}^c \quad [\%], \tag{5.6}$$

$N^{c}_{correct}$ represents number of correctly classified inputs of class c from the set of classes C.

A suitable and frequently used method for determining the classification accuracy of each class is confusion matrix [74]. Field of confusion matrix is described in Fig. 5.2 below.

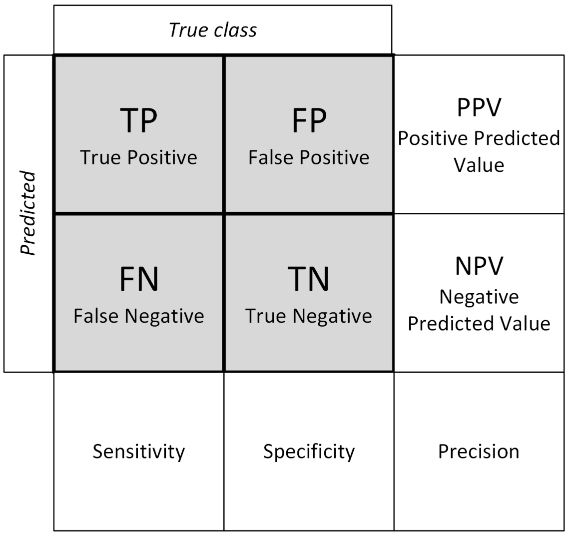


Fig. 5.2: Confusion matrix - description of fields.

Performance measures calculated from mentioned matrix [75]:

- Accuracy

$$A_{CM} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5.7}$$

- Sensitivity (recall, true positive rate)

$$Sensitivity = \frac{TP}{TP + FN}. \tag{5.8}$$

- Specificity (true negative rate)

$$Specificity = \frac{TN}{TN + FP}. \tag{5.9}$$

- Precision (predicted positive value)

$$PPV = \frac{TP}{TP + FP}. \tag{5.10}$$

- False positive rate

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity. \tag{5.11}$$

- False negative rate

$$FPR = \frac{FN}{FN + TP} = 1 - Sensitivity. \tag{5.12}$$

Another very powerful tool for the evaluation and adjustment of the classifier is a graphic visualization by Reciever Operating Characteristic (ROC) or ROC curve. This curve shows the relationship between sensitivity 5.8 (recall, TPR) and FPR 5.11. ROC is providing a visual representation of the relative relationship between pairs of benefits TP and prices FP of the binary classifier, as shown Fig 5.3. Area Under Curve (AUC) is used to evaluate the effectiveness of numerical classifier.

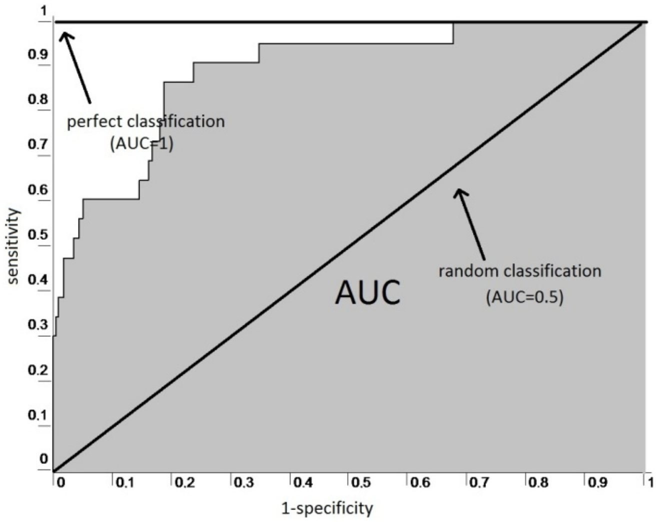


Fig. 5.3: Example of ROC. [76]

6 THE FIRST DRAFT OF CLASSIFICATION SYSTEM AND EVALUATION ON REFERENCE DATABASE

This chapter is devoted to comparing the accuracy of the classification methods described in Sec. 5and proposal of the new approach for speech emotion recognition system. Emotions recordings used in this experiment come from well-known Berlin Database of Emotional Speech - Emo-DB (BerlinDB) [77].

6.1 BerlinDB

For this system was used Berlin database emotions that belong to the group of acted emotions. The database is created by a voice of professional actors without background noise (in the echo-free chamber). Recordings are located in seven emotional states and were recorded by five men and five women. Actors simulate seven emotional states on ten sentences in the German language. These recordings were presented to twenty students for the independent evaluation of specific emotions. The recordings with highest confuse evaluation were removed from the database. The final database contains more than 530 recordings. Given the complexity of SER task, five emotional states formed the study population (anger, fear, happiness, neutral, and sadness).

6.1.1 Feature extraction and selection

Feature set consists of 34 LLDs features and their delta coefficients for all segments and 4 LLDs with delta coefficients for voiced regions only. Functionals are computed from each LLDs. Detailed list of each feature is subscribed in Sec. 4.5. Table 6.1 shows the number of coefficients.

Tab. 6.1: Number of features from different categories.

extraction area	all regions	voiced regions only
LLDs	34 + 34△	4 + 4△
functionals	21	19
Total	1582	

The number of features is significant at first sight. Therefore was feature vector selected with PCA method (Sec. 4.6). Reduced vector do not yield the expected results. On the one hand, a number of features rapidly decreased but also drastically decreased the accuracy of all classifiers. The conclusion of this analysis is a decision not to apply the method for feature selection due to the main objective of research which is highest possible accuracy.

6.2 Selection of classifier

The previous experience led to the selection of three classification methods described in Sec. 5. The feature vector is defined in Sec. 4.5 is the input data for the classifier. The vector is divided into training and test subsets with cross-validation method [78].

6.2.1 Emotion recognition - 5 emotions

The objective of this experiment is the selection of the most accurate classifier considering to the recognition of five emotional states. The following tables show the precision k-NN, SVM and Feed-Forward Back Propagation Neural Network (FFBP-NN). Data and classifiers have been set as follows:

- **Data:** 5 emotional state from BerlinDB, aprox. 404 recordings (15% for testing), 1582 features.
- **k-NN** 10 neigrbours, euclidean distance metric, squared inverse distance weight.
- **SVM** kernel function: cubic, automatic kernel scale, one-vs-one multiclass method.
- **FFBP-NN** number of neurons: 10n in hidden layer, 5n in output layer

Results:

Tab. 6.2: k-NN confusion matrix for 5 emotional states with **77%** precision.

Predicted	*True class*				
Anger	**70**	0	25	5	0
Fear	0	**91**	9	0	0
Happiness	13	12	**75**	0	0
Neutral	0	11	16	**68**	5
Sadness	0	12	0	6	**82**
[%]	Anger	Fear	Happiness	Neutral	Sadness

Tab. 6.3: SVM confusion matrix for 5 emotional states with **80%** precision.

Predicted	*True class*				
Anger	**71**	6	18	10	0
Fear	0	**80**	20	0	0
Happiness	25	8	**67**	0	0
Neutral	0	6	6	**81**	0
Sadness	0	1	0	0	**99**
[%]	Anger	Fear	Happiness	Neutral	Sadness

Presented tables describe the precision for different types of classifiers that have been trained and tested with the above-described data set. Studied group contains five emotions (classes). Each classifier recognized emotional states with similar precision, in other words, there is not one emotion more significant from the others. Overall average accuracy is 77% for k-NN, 80% and 81% for FFBP-NN. Previous studies promised relatively high precision

Tab. 6.4: FFBP-NN confusion matrix for 5 emotional states with **81%** precision.

Predicted	*True class*				
Anger	**72**	6	22	0	0
Fear	0	**88**	13	0	0
Happiness	25	8	**75**	0	0
Neutral	0	0	14	**69**	18
Sadness	0	1	0	0	**99**
[%]	Anger	Fear	Happiness	Neutral	Sadness

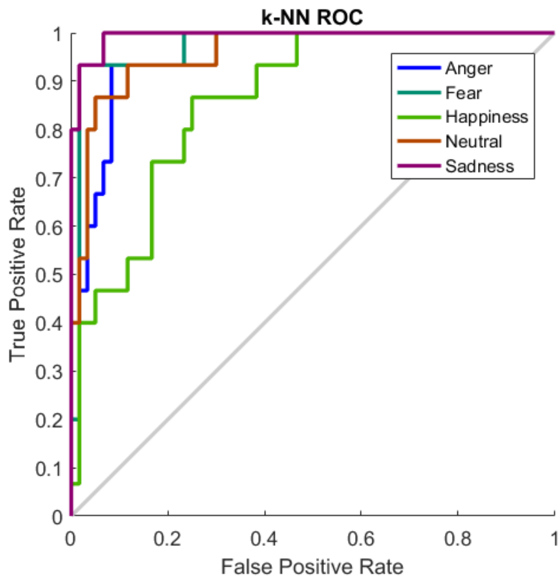


Fig. 6.1: ROC for k-NN classifier for 5 emotional state of BerlinDB.

of chosen classification methods [Par15], especially for the results from Tab. 6.4 achieved by FFBP-NN. Figures 6.1, 6.2 and 6.3 show ROCs for five emotional states of k-NN, SVM and FFBP-NN classifier, where the best result are represented by curves moved to the top left corner.

6.2.2 Cross-emotion recognition

The second part of the experiment aims to verify the cross-emotional performance. The classification model is represented by emotion couple. The result is mutual recognition ability of both coupled emotions. The number of trained models is 10 (all combinations). The model is created by FFBP-NN classifier, since its highest precision. Cross-emotion precision of each couple are listed in Tab. 6.5.

At first glance, it is clear that classifiers achieved a nearly perfect level of precision. The results promise a high percentage but in the following experiment of parallel cross-emotion model is proven otherwise.

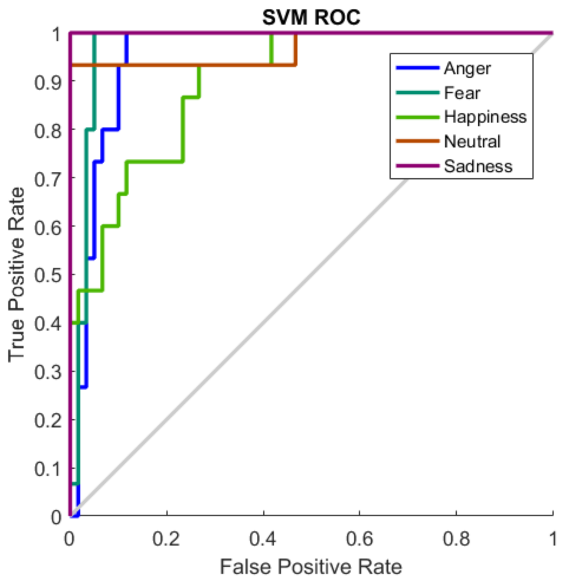


Fig. 6.2: ROC for SVM classifier for 5 emotional state of BerlinDB.

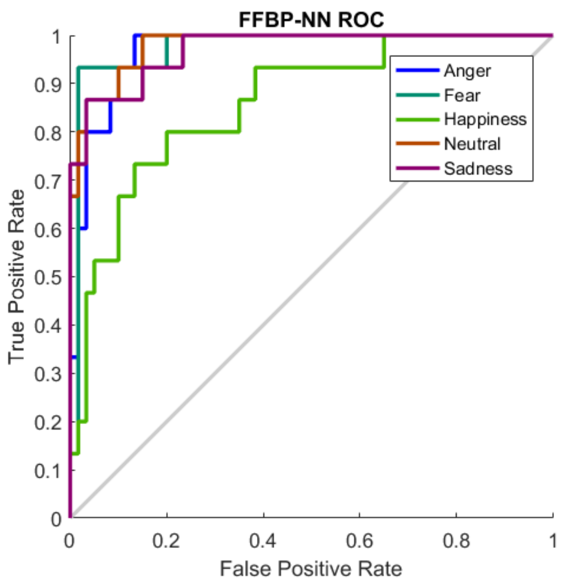


Fig. 6.3: ROC for FFBP-NN classifier for 5 emotional state of BerlinDB.

Tab. 6.5: Precision of cross-emotion recognition for each presented couple. (FFBP-NN classifier)

[%]	Anger	Fear	Happiness	Neutral	Sadness
Anger	-	99	98	100	100
Fear	99	-	97	97	100
Happiness	94	95	-	100	100
Normal	100	98	100	-	99
Sadness	100	98	100	98	-

6.3 Parallel emotion couple recognition system - first classifier proposal

The idea was to use the findings from the experiment in Sec. 6.2.2. The effort was to create a system using the parallel fusion of multiple models. The system consists of 10 classifiers trained on pairwise combinations of emotional states. One classifier is trained by features of the emotional couple. The precision of the results from Tab. 6.5 is the reason to choose FFBP-NN for each of the ten models. The design of described system is shown in Fig. 6.4.

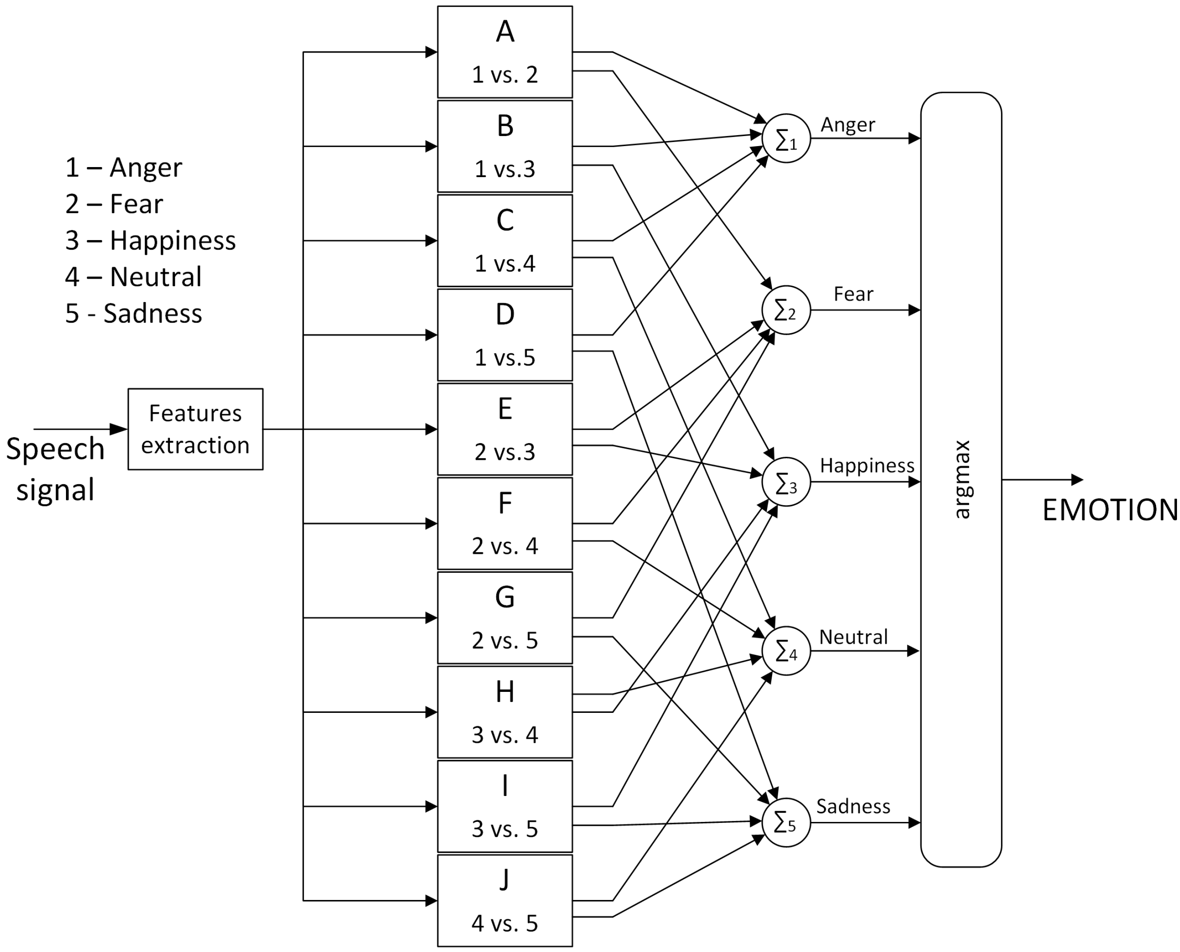


Fig. 6.4: Parallel cross-emotion recognition system for 5 emotions. System contains 10 model for emotion couples.

The feature vector of studied sample is forwarded to all classifiers. One emotion score is calculated separately and compared with each other. 4 classifiers, which include a given emotion (for example score for emotion 3 - happiness is provided by results from classifiers B,E,H,I) create the score for the final decision. Final determination rule about the classes c is formulated by Eq. 6.1, where $P^{D(k)}(\omega_j|x_F)$ is the posterior density distribution of the emotion category ω_j for the feature vector x_F from the classifier $D(k)$.

$$c = \text{argmax} \sum_{k=1}^{K} P^{D(k)}(\omega_j|x_F). \tag{6.1}$$

The decision rule is made by argmax function from the sum of posterior density distributions.

$$c = \begin{cases} \textbf{class} & \qquad\qquad\qquad\qquad \textbf{score} \\ 1 \quad for & \left(P^{D(A)}(\omega_j|x_F) + P^{D(B)}(\omega_j|x_F) + P^{D(C)}(\omega_j|x_F) + P^{D(D)}(\omega_j|x_F)\right) \\ 2 \quad for & \left(P^{D(A)}(\omega_j|x_F) + P^{D(E)}(\omega_j|x_F) + P^{D(F)}(\omega_j|x_F) + P^{D(G)}(\omega_j|x_F)\right) \\ 3 \quad for & \left(P^{D(B)}(\omega_j|x_F) + P^{D(E)}(\omega_j|x_F) + P^{D(H)}(\omega_j|x_F) + P^{D(I)}(\omega_j|x_F)\right) \\ 4 \quad for & \left(P^{D(C)}(\omega_j|x_F) + P^{D(F)}(\omega_j|x_F) + P^{D(H)}(\omega_j|x_F) + P^{D(J)}(\omega_j|x_F)\right) \\ 5 \quad for & \left(P^{D(D)}(\omega_j|x_F) + P^{D(G)}(\omega_j|x_F) + P^{D(I)}(\omega_j|x_F) + P^{D(J)}(\omega_j|x_F)\right) \end{cases} \tag{6.2}$$

Table 6.6 and Fig. 6.5 show the final decision of system (according to sum of classifiers). The score is obtained from the testing with the test set of recordings. The results show that the design is not entirely suitable. An error occurred in the emotions of fear and happiness when it was incorrectly classified anger. The error results from the principle of individual sub-classifiers. Each sub-classifier selects one class as a result of the classification. This means that does not work with a universal background model, and always selects one of its internal emotional states.

Tab. 6.6: Score of classified emotion (top of table) for true emotion testing data (left of table). [%]

True Class	Anger	Fear	Happiness	Neutral	Sadness
Anger	**99**	50	75	25	0
Fear	**83**	73	69	18	7
Happiness	**90**	50	85	25	0
Neutral	16	81	49	**83**	19
Sadness	5	63	20	62	**100**

For example, recordings of happiness and fear will be classified as anger, because score Σ_1 reached the highest value, which is highlighted in Tab. 6.6 and showed in Fig. 6.5, where purple column represents Σ_1 score for anger decision. It must be said that this proposal is not appropriate and will not be further developed in this work. However, presented proposal is not used in this work but it is a challenge for the further research. The idea to use pair models could be used in other applications. One of application could be to speaker identification. One sub-classifier is trained on voice searched for a man versus background model.

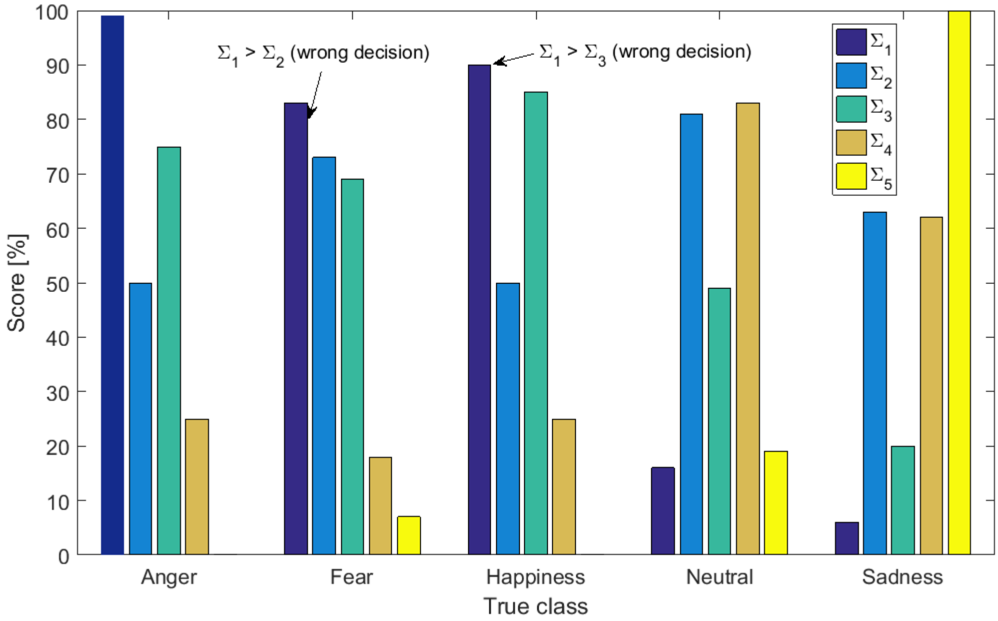


Fig. 6.5: Score of classified emotion of parallel cross-emotion recognition system.

7 EXPERIMENTAL EVALUATION OF CLASSIFIERS ON NEW CREATED DATABASES

This chapter focuses on the speech emotion recognition from the Czech language. The aim was to create a database of emotional recordings from Czech speech and design a SER system for classifying emotional states from this database.

7.1 Czech speech database - emoDBova

In general, the development of speech corpus is divided into three phases. It is necessary to define the purpose of the database clearly, accordingly determine the content of the speech corpus, and define the number and type of speakers and recording conditions.

The effort was the creation of a database for the purpose of training and testing SER system. The system should be content and gender independent. The database (emoDBova) was created by the Department of Telecommunications, VSB TU-Ostrava. Speech source has become show from the Czech radio station. Show content consists of phone calls where the moderator tried to trap called people. The result was quite a strong emotional stimuli from a stressful situation. Groups of emotions are represented by anger, happiness, sadness and normal. The disadvantage of this source is the absence of fear because it has been stimulated very weak. Fear absence was confirmed by the result of subjective evaluation also [Uhr14].

Recordings were freely available on the official website of the radio station and the youtube.com website, where the only audio part was taken. Recordings respect the following characteristics:

- Duration of record from two to six seconds.
- Recording should not contain environmental noise.
- recording has to include human speech in the form of few words or a full sentence, not only interjection.

As an output format for database samples have been used waveform audio file. Audio files parameters are:

- 16-bit PCM
- mono channel
- $f_s = 16$ kHz
- bitrate 128 kbps

For better orientation and work with the database, the name of the recording carries information about emotion, gender and the unique identifier of the recording. The database contains a total of 439 recordings [Uhr16].

7.2 Subjective evaluation

Recording of the database was evaluated by subjects (students) in the age range from 18 to 26 years. Approximately 10 subjects rated each recording. Web environment for subjective evaluation are shown in Fig. A.1 located in Appendix A. Recording of different emotion, gender and different accuracy of subjective evaluation can be exported from the emoDBova. The example of exported recordings from the database are shown in Tab. 7.1 below.

Tab. 7.1: Example of exported database.

ref_id	**value_of_veracity** [%]	**sp_gender**	**final_emotion**
s016f	46.67	female	Neutral
h011m	85.71	male	Happiness
a091f	87.50	female	Anger
a115m	100.00	male	Anger
n066m	62.50	male	Neutral
		number of samples: 5	

Column *ref_id* is the label of an audio file, where the first letter means emotion and last one is gender. The *value_of_veracity* defines subjective evaluation results of listeners (students). The first line of the table is an example when the recording is evaluated as neutral despite the fact that was marked as sadness during the creation of the database (human factor error).

Table 7.2 describes the number of each emotional states recordings and levels of evaluation (veracity).

Tab. 7.2: Database veracity and quantity.

Emotion	**Veracity** [%]	**Quantity**	**Gender**
Anger	93.75	128	equal
Happiness	51	76	equal
Neutral	68	164	equal
Sadness	72	71	equal

The table shown reduced significance recordings marked with happiness. Attached Fig. A.2 and A.3 show environment for extraction settings from database and Fig. A.4 shows text list of extracted records. The next section will describe the results of classification.

7.3 Classification of emoDBova

Just as in Sec. 6.2.1, same SER classifiers are trained for the Czech language. Training and test data are extracted from a database which is described in the previous section. Feature vector contains the same parameters as in the BerlinDB case. k-NN, SVM and FFBP-NN is selected for the classifier role.

7.3.1 Results

Tables 7.3, 7.4 and 7.5 show achieved precision of individual classifiers for recordings from emoDBova. The lowest ability achieves k-NN and SVM for the emotional state of neutral. A remarkable finding is that recordings marked as neutral were often classified as *rest in set* (40% of not-neutral from k-NN, 55% of not-neutral from SVM). One reason may be the lack of significant parameters for the separation of emotional states of neutral and other emotions. A Very suitable classifier for presented task seems k-NN, which achieved the best average result with 76% precision.

Tab. 7.3: k-NN confusion matrix for 4 emotional states from emoDBova with average precision of **76%**.

Predicted	*True class*			
Anger	**87**	0	13	0
Happiness	0	**80**	10	10
Neutral	5	20	**60**	15
Sadness	0	21	0	**79**
[%]	Anger	Happiness	Neutral	Sadness

Tab. 7.4: SVM confusion matrix for 4 emotional states from emoDBova with average precision of **71%**.

Predicted	*True class*			
Anger	**100**	0	0	0
Happiness	0	**69**	8	23
Neutral	42	7	**45**	6
Sadness	0	29	0	**71**
[%]	Anger	Happiness	Neutral	Sadness

A major drawback of emoDBova is an absence of the anger emotion. For this reason, it is not possible to compare the results with the previous experiment from Sec. 6.2.1. On the other hand, it is certainly possible to express the expected finding. In the newly formed emoDBova does not constitute such significant differences between emotional states as in BerlinDB. Despite the smaller number of classified classes was achieved lower accuracy.

Tab. 7.5: FFBP-NN confusion matrix for 4 emotional states from emoDBova with precision of **68%**.

Predicted	*True class*			
Anger	**93**	0	0	7
Happiness	0	**56**	11	33
Neutral	10	20	**65**	5
Sadness	0	35	6	**59**
[%]	Anger	Happiness	Neutral	Sadness

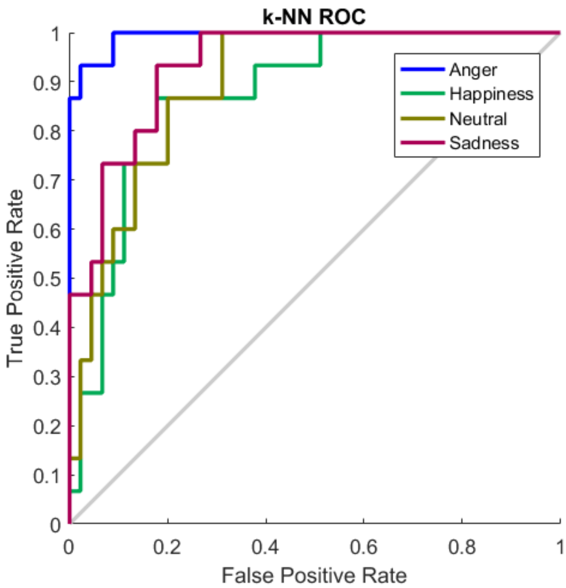


Fig. 7.1: ROC for k-NN classifier for 4 emotional state of emoDBova.

This fact is the precondition for decreasing precision with the same classification dimension as in BerlinDB task.

The result of this experiment confirms mentioned classification accuracy dependence on the database type. The precision of classification methods on BerliDB was significantly higher than in this case. Reduced precision was expected whereas emoDBova is not a database from a recording studio but consist of normal telephone calls in radio shows. Classifiers precision for BerlinDB was highest for FFBP-NN and the lowest k-NN. In the case of emoDBova, the results were just the opposite (best for k-NN). These results are clearly visualized by ROC curves in Fig. 7.1, 7.2 and 7.3. The effectiveness of this method shows a bend of curves to the upper left corner.

Comparison of the subjective evaluation and the classification results are shown in Fig. 7.4. For classes anger, happiness and sadness are classification accuracy higher than subjective evaluation (except FFBP-NN for happiness and sadness). A subgroup of neutral recordings was classified with worst precision. From the Tab. 7.3, 7.4 and 7.5 is clear that

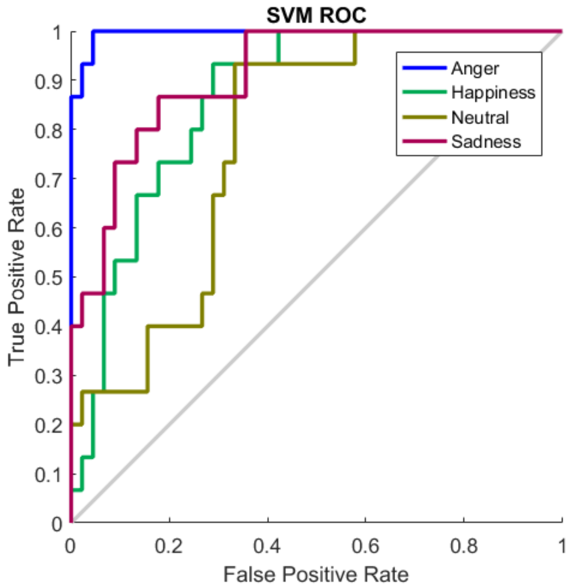


Fig. 7.2: ROC for SVM classifier for 4 emotional state of emoDBova.

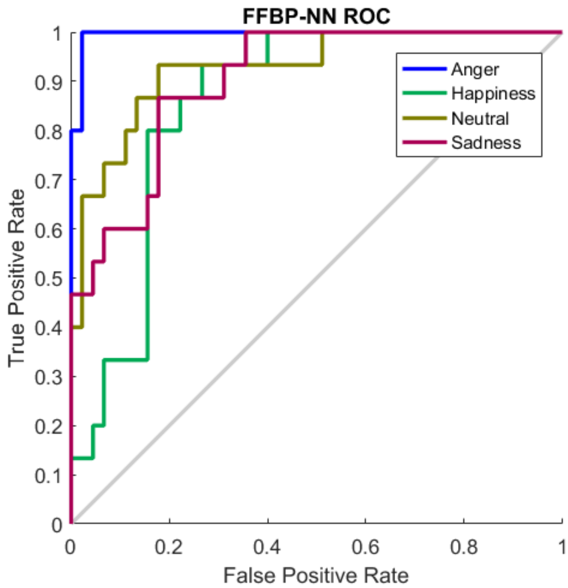


Fig. 7.3: ROC for FFBP-NN classifier for 4 emotional state of emoDBova.

the neutral was often misclassified as happiness and anger. For this reason, the recording with lower emotional significance will be replaced and again subjectively evaluated to preserve and enhance the quality of the database. As already mentioned several times, the quality of the database directly affect the classification accuracy. Therefore, the key step is selecting correct data for classifier training set. It would adversely affect their accuracy and precision.

The chapter 8 describes the design proposal of a classifier for emoDBova. The results

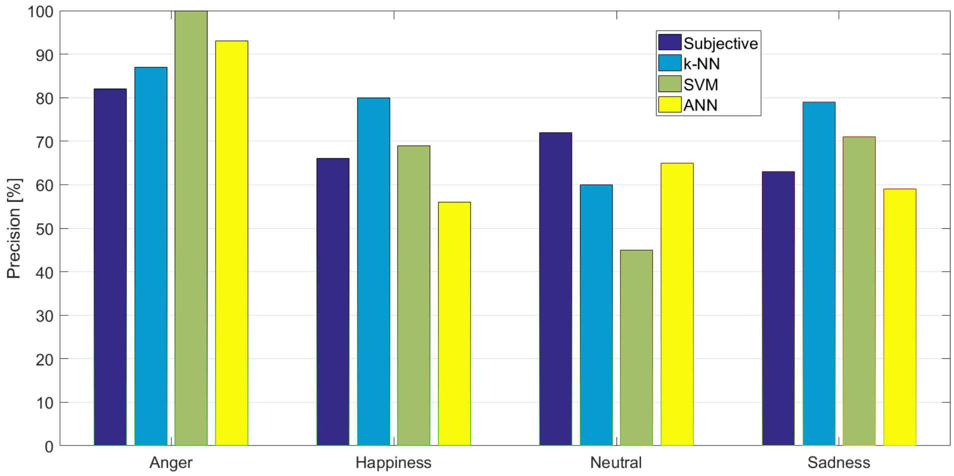


Fig. 7.4: Comparison of subjective evaluation and k-NN, SVM and FFBP-NN classification precision fro emoDBova.

are compared with this and the previous experiment.

7.4 Additional databases

As part of the Ph.D. study were created other databases that focused on SER. One database contains recordings of the Integrated Rescue System (IRS) of 112 emergency line. It was designed to detect the stress from the human voice. The second database consists of audio recordings from Czech language movies. The source of this database is acted speech (not spontaneous), which means that emotions are simulated. The results from this database are not detailed presented to avoid unnecessary theses inflating.

7.4.1 Czech speech database - 112DB

Firstly, it should be noted that all source data has been anonymised given the subject to the Law on Personal Data Protection.

The number of database recordings is small but on the other hand, has a relatively high value for the design of a speech stress detection system.

The database includes 31 recordings of calls between callers in need and emergency call center agents. Recording length varies in the range of 30 seconds to 10 minutes. The content of the call is in most cases from an unfortunate event (car accident, death, violence and other incidents). Content is unsuitable for children and sensitive listeners. Recordings are divided according to the voice of a caller in need, which form the stress samples and recording of the agent's voice (neutral state).

Properties of database:

- 31 recordings
- 2 emotional states, 31 agents for neutral and 31 stressed callers
- Mono, 8kHz, 32-bit float
- without gender information

Speech stress detection system has clearly defined the field of applicability. The system is useful in situations with increased stress stimuli (police, military or fire dept.). From the radio channel can be extracted and recognized the man under stress. Dispatching (commander) can react and adjust the tactical procedure.

Table 7.6 lists the results of k-NN, SVM and FBP-NN classifiers in the role of stress detector. Feature vector consist from the same features as describes Sec. 4.5.

Tab. 7.6: Precision of classifiers on neutral-stress recognition task from 112DB.

[%]	**k-NN**	**SVM**	**FFBP-NN**
Neutral	73	96	96
Stress	95	97	100

The results show the presence of strong emotive stimulated recordings. As in previous experiments, FFBP-NN achieves the best results. The smaller number of samples must be taken into account. For this reason and from previous researches, it can be argued that the accuracy will decrease with the expansion of the database. The experiment confirmed that the database (in this form) is suitable for the testing speech stress detection systems.

7.4.2 Czech speech database - emoMovieDB

The database was created by students within the scope of Multimedia Technologies at Dept. of Telecommunications, Technical University of Ostrava. A source of the database is voices of actors from Czech language movies. Properties of database:

- 680 recordings
- 5 emotional states (anger, fear, happiness, neutral, sadness)
- Stereo, 48kHz, 32-bit float
- gender information included

No recordings were subjectively evaluated. Another disadvantage is the lack of real emotional stimuli. In further research, it will be used, but only after the evaluation and selection of less significant recordings.

8 PROPOSAL OF MULTI-CLASSIFIER SER SYSTEM AND VERIFICATION OF NEW APPROACH

One of the thesis objectives is classifier design for speech emotion recognition. The proposal should be aimed at increasing accuracy of recognition. Data fusion can be performed at three levels: data level fusion, feature level fusion and fusion classifier [79]. There are several approaches to reach higher accuracy, but the most common is a classifier fusion. Multi-Classifier Systems (MCS) focus on the combination of classifiers from heterogeneous or homogeneous modeling backgrounds to give the final decision [80], [81].

8.1 MCS design

The proposal includes a classification method of experiments presented in Sec. 6. Individual classifiers are connected in parallel structure. The input vector of feature extraction is presented to each classifier. Classifier's output is predicted class (winner) or posterior probability. It will represent the input for the last block - fusion. The pipeline of Fig. 8.1 shows a proposal of MCS for emotion recognition.

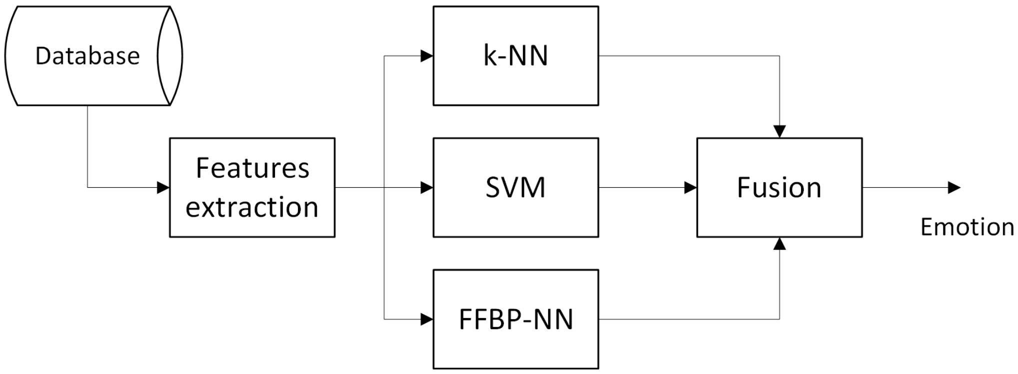


Fig. 8.1: Proposed MCS for emotion recognition based on fusion of three classifiers.

8.2 Fusion of the verification rate

An important aspect in a fusion of the verification rate is score normalization. The role of normalization is the mapping of each classifier score to the same domain (e.g. range 0–1). This condition is fulfilled from the first experiment since the output of the classifier is a hard decision on the classified class or posterior probability in the range mentioned above. The fusion of the verification rate is a combination problem. Scores from the individual classifiers are combined into a scalar score for the final decision. The combination of score expressed as posterior probabilities use If the score is expressed posterior probabilities are used following a combination of methods. For the system design were used the following combination methods for posterior probabilities.

- **Product rule** - this rule is based on the assumption that the feature vectors are statistically independent. The speech signal represented by feature vector x_F is assigned to the resulting class c according to the formula:
$$c = \underset{j}{\operatorname{argmax}} \prod_{k=1}^{K} P^{D(k)}(\omega_j|x_F), \tag{8.1}$$
where $P^{D(k)}(\omega_j|x_F)$ is posterior probability that feature vector x_F belongs to j class. Posterior probability is output of k^{th} classifier of total number of K classifiers.
- **Sum rule** - this rule takes into account that the posterior probability of all classifiers is not very different from the a priori probabilities of each class (count of one class against the other). Vector x_F is assigned to a class c based on the equation below.
$$c = \underset{j}{\operatorname{argmax}} \sum_{k=1}^{K} P^{D(k)}(\omega_j|x_F), \tag{8.2}$$
where the meaning of the variables corresponding to the previous formula.

8.3 Bayes belief integration

The approach mentioned above affects all classifiers as well and does not include errors that were produced by each of them. These errors can be simply described by confusion matrix as
$$PT_k = \begin{pmatrix} n_{11}^{(k)} & \cdots & n_{1(M+1)}^{(k)} \\ \cdots & \cdots & \cdots \\ n_{M1}^{(k)} & \cdots & n_{M(M+1)}^{(k)} \end{pmatrix}, \tag{8.3}$$
where rows are corresponding to true classes $c_1, c_2, \cdots c_M$ and the columns represents the predicted classes e_k by classifier $D(k)$. The values $n_{ij}^{(k)}$ reflect how much of input samples of class c_i (speech passes in the emotional state i) were classified as c_j. From confusion matrix PT_k can be derived belief measure of correctly classified as follows:

$$Bel(x_F \in c_i/e_k(x_F)) = P(x_F \in c_i/e_k = j), \tag{8.4}$$
where $i = 1, \cdots, M$ and $j = 1, \cdots, M+1$ and
$$P(x_F \in c_i/e_k(x_F) = j) = \frac{n_{ij}^{(k)}}{\sum_{i=1}^{M} n_{ij}^{(k)}}. \tag{8.5}$$

The combination of defined belief measure for each classifier form a new belief measure for multiple classification system as follows:

$$Bel(i) = P(x_F \in c_i) \frac{\prod_{k=1}^{K} P(x_F \in c_i/e_k(x_F) = j_k)}{\prod_{k=1}^{K} P(x_F \in c_i)}. \tag{8.6}$$

Probability described in the equation above can be easily calculated from the confusion matrix. Class with the highest $Bel(i)$ is chosen as the final decision of classification system.

8.4 Results verification

The above-described fusion methods have been applied, and the system has been tested. Feature vector and parameters of the classifiers were kept the same, due to the possibility to compare the precision of previous experiments. Below shown tables represent the confusion matrices for samples (recordings) of BerlinDB and emoDBova.

Tab. 8.1: Confusion matrix for designed system with **sum rule fusion**. Values describe precision of system on **BerlinDB** samples. Average precision is **83%**.

Predicted	*True class*				
Anger	**71**	4	20	5	0
Fear	0	**87**	13	0	0
Happiness	20	9	**71**	0	0
Neutral	0	0	11	**82**	6
Sadness	0	0	0	0	**100**
[%]	Anger	Fear	Happiness	Neutral	Sadness

Table with confusion matrix for product rule fusion is not presented here because of minimal differences with **sum rule fusion**. The system with product rule fusion reaches 82% average precision.

Tab. 8.2: Confusion matrix for designed system with **bayes belief fusion**. Values describe precision of system on **BerlinDB** samples. Average precision is **85%**.

Predicted	*True class*				
Anger	**70**	5	20	4	0
Fear	0	**88**	12	0	0
Happiness	22	0	**78**	0	0
Neutral	0	0	12	**88**	0
Sadness	0	0	0	0	**100**
[%]	Anger	Fear	Happiness	Neutral	Sadness

Section 6.2 presents the results of the classification of emotional states using the k-NN, SVM and FFBP methods separately. The data come from BerlinDB and precision is shown in Tab. 6.2, 6.3 and 6.4.

The fusion of these classifiers has similar results. The emotional state of sadness achieved best results after classifiers fusion same like in the experiment of Sec. 6.2. Therefore, it can be argued that the selected feature vector contains enough significant parameters to distinguish sadness from other emotional states (Tab. 8.1 and 8.2). Exactly the opposite proposition arises with analyzing the results for anger and happiness. In all classification approaches, there was a high percentage of mutual misclassification of anger and happiness.

Fusion methods can be evaluated positively. The overall accuracy after applying the sum rule and product rule achieved just minor impact. The precision of 83% is only a 2% increase over separate FFBP-NN has reached 81%. In such cases, it is necessary to compare the benefits vs. price of the solution. In this case, it is the 81% with the single classifier vs. 2% accuracy increasing with the complexity of three classification methods. From the other side, there is an 85% precision with bayes belief integration, where it is worth to consider computing demands.

Tab. 8.3: Confusion matrix for designed system with **sum rule fusion**. Values describe precision of system on **emoDBova** samples. Average precision is **74%**.

Predicted	*True class*			
Anger	**93**	0	0	7
Happiness	0	**68**	0	32
Neutral	9	12	**73**	6
Sadness	0	31	5	**64**
[%]	Anger	Happiness	Neutral	Sadness

As with BerlinDB, system with product rule fusion on emoDBova database achieve almost the same results as Tab: 8.3.

Tab. 8.4: Confusion matrix for designed system with **bayes belief fusion**. Values describe precision of system on **emoDBova** samples. Average precision is **78%**.

Predicted	*True class*			
Anger	**90**	0	3	7
Happiness	0	**76**	0	24
Neutral	0	17	**78**	5
Sadness	0	29	4	**67**
[%]	Anger	Happiness	Neutral	Sadness

A similar situation arose with the classification of samples from emoDBova. Sum rule and the product rule did not reach the expected benefits. Moreover, this types of fusion do not exceed the precision of the simple k-NN classifier (k-NN 76% vs. sum rule fusion of k-NN, SVM and FFBP-NN 74%). Bayes belief integration again achieved better results. This fusion increased the precision from 76% to 78% percent. The ROC characteristics of proposed system are shown in Fig. 8.4 and Fig. 8.4.

Regarding individual emotional states, there is misclassification between happiness and sadness. Emotional state anger achieved the best precision.

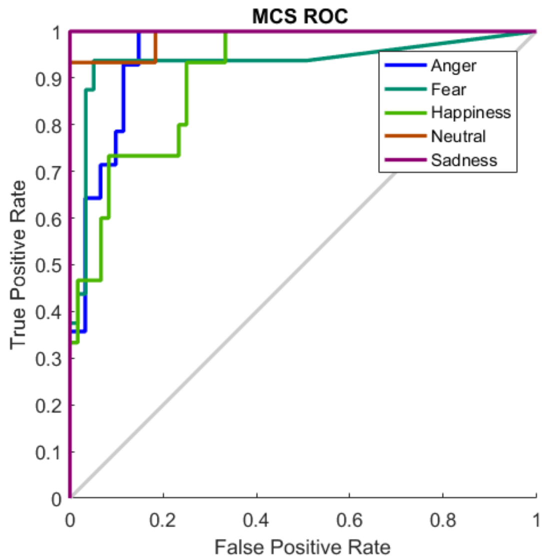


Fig. 8.2: ROC for designed system on BerlinDB.

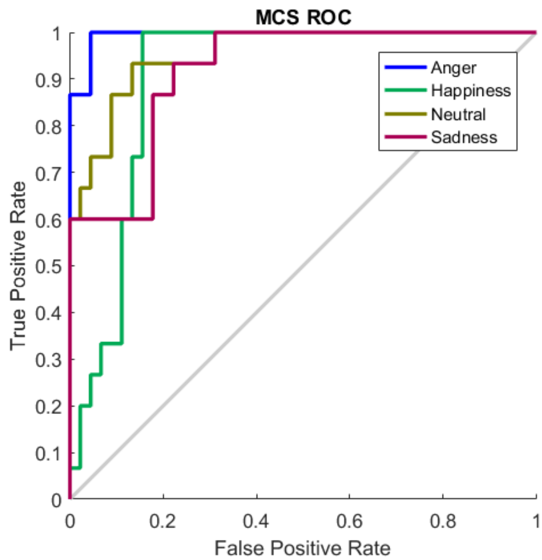


Fig. 8.3: ROC for designed system on emoDBova.

8.5 Summary and comparison

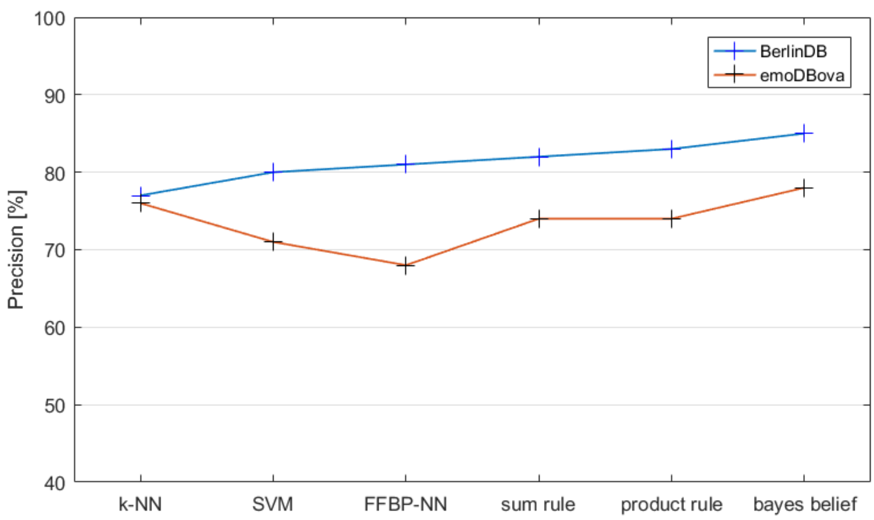
All research described in this work towards to designing a system for speech emotion recognition, especially for Czech speech. To achieve this aim, it was necessary to create an appropriate database for this purpose. Czech emotional database emoDBova was created out of spontaneous speech for system training and testing [Uhr16]. For comparison of the results was used BerlinDB. The composition of feature vector was also one of the key steps. Based on previous research [Par15], [Par16] and [Par14] were selected parameters listed in Sec. 4.5. From the contours of features was calculated statistical values defined as functionals. The choice of classification methods is also based on the results from published research [Par15].

Table 8.5 shows the results obtained from the system design. Precision k-NN classifiers, SVM, and FFBP-NN reached 77-81 percent for BerlinDB, confirm the assumptions about chosen classifiers. In the case of newly created emoDBova classifiers achieved lower precision (68-76%). Due to the more accurate classification of the emotional state, I decided to take a more comprehensive multi-classification system. MCS is based on the fusion of three verified classifiers. Three types of fusion have been evaluated. Bayes belief MCS reached the best result on emoDBova recordings (). The overall performance of the experiments are shown in Fig. 8.4 7.2

Tab. 8.5: Precision of presented experiments.

Database	k-NN	SVM	FFBP-NN	Sum rule	Product rule	Bayes Belief
BerlinDB (5 emo.)	77	80	81	82	83	85
emoDBova (4 emo.)	76	71	68	74	74	78
112DB (2 emo.)	84	96	98	-	-	-

Fig. 8.4: Precision comparison of evaluated classifiers and proposed system on BerlinDB and emoDBova.



8.5.1 Comparison with related research

A typical feature of all the presented research is the comparison of results (mainly accuracy) with previous proposals. Not only my opinion is that such a comparison is not entirely appropriate. It can be argued that there are not two experiments which used the same feature vectors, classification methods, but especially the type of database. These design attributes must be included in the comparison. Table 8.6 lists several relevant systems that are at least in part like this work.

Tab. 8.6: Comparison of related works with design attributes.

Source	Emotions	Database	Features	Classification	Achieved results
Pappas 2015 [82]	2 emotions anger vs. rest (neutral)	call center (Greek)	MFCC, E, ZCR, F0, + functionals	Logistic regresion	70% (anger detection)
Lugger 2007 [35]	6 emotions happiness, bored, neutral, sad, angry, anxious	BerlinDB (German)	MFCC, LFPC, VQP, + functionals	GMM	67–74%
Vaudable 2012 [83]	3 classes negative, neutral, positive emotions	call center (French)	MFCC, F0, formants, + functionals	SVM	80% (negative vs. positive emotions)
Atassi 2012 [84]	5 emotions anger, happiness, neutral, sadness, surprise	call center (Slavic)	MFCC, HFCC, PLP, and others, + functionals	Two phase classification	74%
Own proposed system	4 emotions anger, happiness, neutral, sadness	emoDBova calls from radio show (Czech)	Listed in Sec. 4.5	Bayes belief fusion of k-NN, SVM, and FFBP-NN	78%

Pappas in [82] presented research that detects the occurrence of the emotional state of anger, it means that the other emotional states formed a background model and amounted to 70% percent (as Vaudable 80% in determining the polarity [83]). Really comprehensive solution chose Lugger presented in [35], who first classified two subgroups of emotional states and then determined the final decision. The results were achieved on BerlinDB. Atassi in [84] presented the results of his proposal with attributes closest to my work. The first classifier selects two most probable emotional states, the second stage is a classification based on cross emotion (model trained only for a specific emotional couple.) This proposal achieved an average 74% accuracy. The system presented in this work amounted the highest 78% for database emoDBova with classes and 85% for BerlinDB with 5 classes. However, it has already been mentioned that this comparison is only approximate.

9 EXPERIMENTAL IMPLEMENTATION OF PROPOSAL SYSTEM IN SECURED COMMUNICATION INFRASTRUCTURE

Created SER system is deployed as an additional service within the project TACR with PID: TF01000091. Name of the project is Security of Mobile Devices and Communication. The project is filed under program TF - The program of promoting cooperation in applied research and experimental development through joint projects of technology innovation agencies DELTA. The solution period of the project is between 01/2015–12/2017.

The aim of the project is to create solutions that will support secure communications between mobile devices within the organization or network. Goals of project are divided into:

1. Voice services: provision secure conversation with one or more users.
2. Text services: the creation of a secure conversation with one or more users.
3. Safe file sharing between users. The possibility of sending files to users.
4. Access to secure file storage.
5. Access to the list of confidential contacts organization.
6. Access to the system, which generates access keys to the various subsystems of the organization.
7. Localization of device.

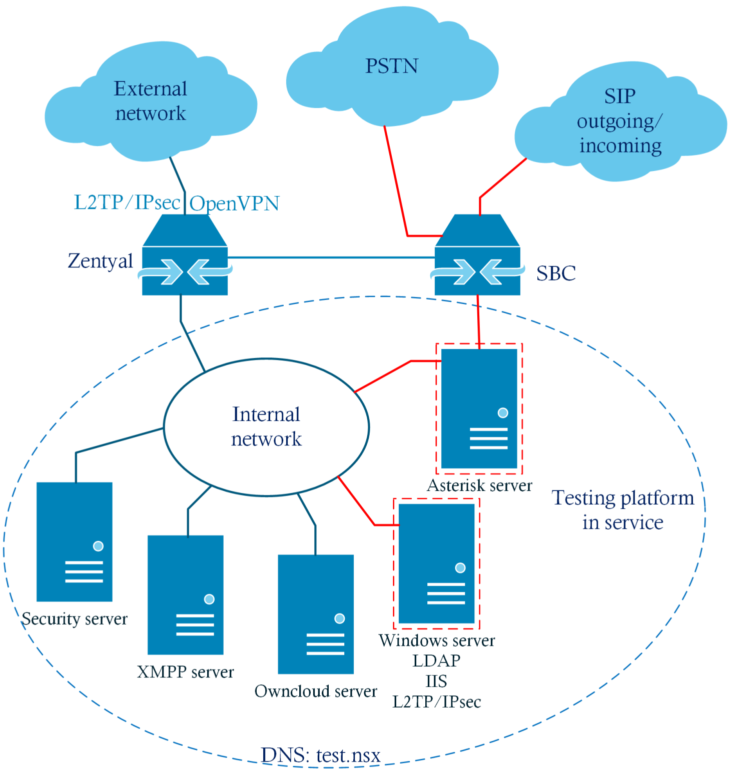


Fig. 9.1: Secured communication system structure for mobile devices.

Project deals with design and implementation of secure communication solutions for multimedia services. The solution will be implemented and used by the police department. Proposed SER system falls under the first of the goals. This means that the SER system analyzes the voice of phone call for the purpose of recognizing emotions of the speaker.

The structure of the communication system is shown in Fig. 9.1. Asterisk server serves as a software PBX for voice services. Dialed calls are recorded directly to the Asterisk server as two separate recordings (caller and called party). These recordings are sent to the Windows server where operate the proposed SER system.

Placed call is immediately processed by the system. OpenSMILE task is the creation of feature extraction and feature vector. This vector is the input for multi-classification system. The caller is associated with the emotional state, and the result is reported to the administrator or other services as shown Fig. 9.2.

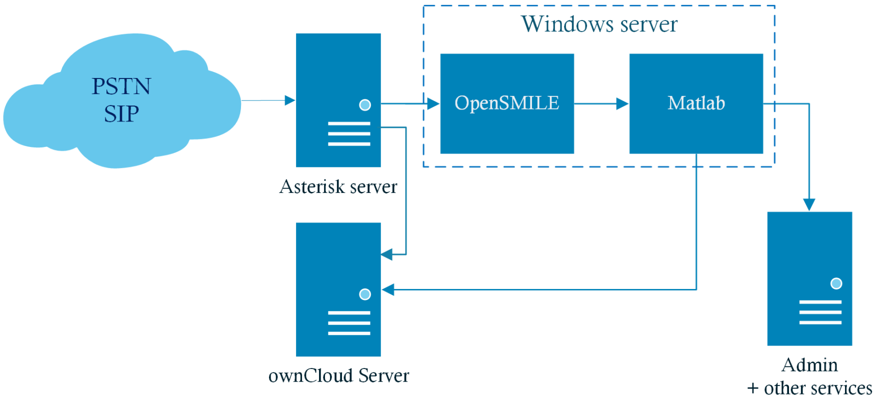


Fig. 9.2: SER system implementation pipeline.

Another not less important benefit of this analysis is the ability to create a new database of emotional recordings since the recorded speech with the classified emotional state is stored in the ownCloud server. The database will record various attributes associated with the call (session) and classified emotions. An example of recordings list extracted from the database is presented in the following table.

Tab. 9.1: Example of database extracted list of records with session attributes and classified emotions.

Time	**Date**	**Session identifier**	**Direction**	**Emotion**
12:30:06	10.11.2016	Alice and Bob	in	3 - Neutral
12:30:06	10.11.2016	Alice and Bob	out	2 - Happiness
00:33:48	09.11.2016	External caller and Bob	in	1 - Anger
00:33:48	09.11.2016	External caller and Bob	out	3 - Neutral
...	...	...	...	...

The record contains a timestamp, session IDs, direction and classified emotional state. The direction field defines who the participants of the session have been uploaded. In other words, the call of one session is divided into two records. One belongs to the caller (in) and the second to an invited participant. Evaluated recordings are stored into the database with followed label:

```
123006_10112016_alice123%bob234_out_2.wav
```

which represent the voice of Bob from session between Alice and Bob recorded in 12:30:06 10^{th} of November 2016 and the record was classified to emotional state happiness.

Implemented SER system will serve to the recognition of the emotional state, but its primary purpose is stress detection from police units voice or calling citizens in need. Sequentially created database will be used to further system retraining. The process of system re-training/re-modeling delivers precise adjustment of classification for a concrete environment.

10 CONCLUSION

The current technological trend emphasizes the simplification and automation of human-machine interaction. The most natural way of information exchange for a man is speech. Human speech contains more information than just content. It is the emotional state, which defines the mental state of man, and even change the apparently clear speech content. This and other uses of emotional states are the reason why speech emotion recognition is a hot topic for many research teams in the field of speech processing. The proof is many relevant publications on this topic in the last decade. Most of the results are difficult to compare because the approach of mentioned proposes often analyzes different speech sources and different databases. Only a few research publications deal with the Czech language speech. These and other reasons are behind the motivation to create this work. Aims of this work were dedicated to the design and implementation of speech emotion recognition system for spontaneous speech in the Czech language.

An important part of the SER system design is a selection of features, which will be the most significant for emotion recognition. For this purpose, 72 LLDs features were extracted, and 21 functionals were derived from LLDs contours. For all feature extraction operations was used OpenSMILE extraction tool. The final feature vector for a single recording was formed in 1582 features. Part of the experiment was the feature selection PCA, but after its application rapidly decreased accuracy and therefore was feature vector used in full length without PCA. The mentioned and following operations have been programmed in Matlab environment. Among the many offered classification methods and after previous research with classifiers the k-NN, SVM and FFBP-NN were selected [Par15]. Their accuracy was verified on the well-known BerlinDB. Feature vectors were extracted from the recordings of five emotional states (anger, fear, happiness, sadness and neutral). At first, the classification accuracy was evaluated on the ability to recognize emotional state of all five classes. The k-NN has reached 77%, SVM 80% and FFBP achieved the highest 81% precision. For FFBP-NN has also been evaluated cross-emotion recognition precision. Percentage of mutual recognition of emotional states (one-by-one) achieved from 94 to 100%. These high values have been a precondition for a first draft of the classification system.

Parallel cross-emotion recognition system was the first draft due to mentioned one-by-one recognition accuracy. The system contained 10 FFBP-NN models for all combinations of emotional couples. The score was determined by the summation rule of models output. One emotion score was extracted from probability density functions achieved from models trained by interest emotion. Unfortunately, the results do not reach the expected preconditions. This proposal often misclassified speech recordings marked in fear and happiness emotion as anger emotion. Therefore, this proposal was rejected and was not further analyzed.

One objective of this dissertation was the creation of new Czech emotional database.

As assignment form indicates, the database was used for training and testing proposed SER system. Therefore, it was important to find adequate speech sources to create such a database. The main prerequisite for the creation of the high-quality database is a spontaneous emotional speech. This kind of speech means recordings of real conversations. The recordings have been systematically edited from the radio broadcast shows, which had a relatively rich content of emotions. The total number of database recordings is 439 in five emotional states (anger, fear, happiness, neutral and sadness). Recordings were exposed to subjective evaluation by students of Department of Telecommunications where the database was called emoDBova. The result of the subjective evaluation is the insufficient emotional significance of recordings marked as fear. Listeners often classified these recordings as other emotional states, in most cases neutral. Only anger, happiness, neutral and sadness was used in rest of work.

Selected classification methods achieve 76% for k-NN, 71% for SVM and 68% for FFBP-NN on emoDBova recordings. The results comparison with BerlinDB classification is not entirely possible given the absence of fear emotion recordings. On the first sight, the increased accuracy could be expected due to the lower number of interested classes. This assumption contradicts differences in sound quality between recordings of BerlinDB and emoDBova. Higher precision in BerlinDB was achieved due to studio-quality recordings, alongside phone channel sound quality in the case of emoDBova.

In addition to emoDBova, the 112DB database has also been created. The source is 112 link of Integrated Rescue System. The number of recordings is small, but on the other hand, voices are strong emotionally stimulated. Recorded speech consist of the phone calls of people in need (car accident, injury, death, domestic violence, and so on). This database was analyzed to use it to stress detection. This means that two classes have been classified, neutral and stress. The precision after FFBP-NN classification achieved 96/100% for neutral/stress recognition.

The aim of the work is also a proposal for a new SER system. This proposal consists of a compilation of multi-classification system based on the fusion of classification methods. The purpose was to achieve increased final classification precision with the available methods. Three types of fusion have been used and verified. Product rule, sum rule, and bayes belief integration were applied, and the accuracy was evaluated on BerlinDB and emoDBova recordings. Given the previous precision of classification methods with the extracted feature vector, the system is composed of the parallel fusion of k-NN, SVM and FFBP-NN and same feature vector. The input of fusion is posterior probabilities of classifiers. The best results were achieved with bayes belief integration fusion. Final classification accuracy reached 85% for five emotional states of BerlinDB and 78% of emoDBova database. Parallel fusion of three classifiers based on bayes belief integration is the final multi-classifier system design.

The proposed system is also part of the infrastructure that is used to secure communications developed within TACR with PID: TF01000091. Infrastructure will serve within secure communications for police units. SER system will analyze telephone calls to recognizing emotional states and for stress detection.

Major contributions of this work are namely:

1. New created and evaluated Czech emotional speech database emoDBova and 112DB for training and testing proposed system and systems developed in further researches.
2. New approach and design of the multi-classifier system for speech emotion recognition.
3. Implementation into the real environment. SER system realizes speech emotion recognition of caller in secured communication system infrastructure.

These points also represent the goals of the dissertation thesis. Mentioned databases, proposed and implemented SER design are unique and have never been used.

BIBLIOGRAPHY

[1] PICARD R. W. *Affective computing*. MIT press, 2000.

[2] STANEK, M. and M. SIGMUND. Psychological Stress Detection in Speech Using Return-to-opening Phase Ratios in Glottis. *Elektronika IR Elektronika*. 2015, vol. 21, no. 5. DOI: 10.5755/j01.eee.21.5.13336. ISSN 1392–1215.

[3] RAMAKRISHNAN, S. and I. M. M. EL EMARY. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*. 2013, vol. 52, iss. 3, pp. 1467–1478. DOI: 10.1007/s11235-011-9624-z. ISSN 1018–4864.

[4] REYES, A., P. ROSSO and D. BUSCALDI. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*. 2012, vol. 74, pp. 1-12. DOI: 10.1016/j.datak.2012.02.005. ISSN 0169–023X.

[5] BANSE, R. and K. R. SCHERER. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*. 1996, vol. 70, no. 3, pp. 614–636. DOI: 10.1037/0022-3514.70.3.614. ISSN 1939-1315.

[6] BRESTER, C., E. SEMENKIN, I. KOVALEV, P. ZELENKOV and M. SIDOROV. Evolutionary feature selection for emotion recognition in multilingual speech analysis. In: *2015 IEEE Congress on Evolutionary Computation (CEC)*. Sendai: IEEE, 2015, pp. 2406–2411. DOI: 10.1109/CEC.2015.7257183. ISBN 978-1-4799-7492-4.

[7] FRAGOPANAGOS, N. and J. G. TAYLOR. Emotion recognition in human–computer interaction, *Neural Networks*. 2005, vol. 18, iss. 4, pp. 389–405, DOI: 10.1016/j.neunet.2005.03.006. ISSN 0893-6080.

[8] LIN, Y. P. et al. EEG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Biomedical Engineering*. 2010, vol. 57, no. 7, pp. 1798–1806, DOI: 10.1109/TBME.2010.2048568. ISSN 1558-2531.

[9] VOZNAK, M., P. PARTILA, M. PENHAKER, T. PETEREK, K. TOMALA, F. REZAC and J. SAFARIK. Emotional state and its impact on voice authentication accuracy. In: *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI (SPIE 8750)*. Baltimore: SPIE, 2013. DOI: 10.1117/12.2015719

[10] SCHERER, K. R. Vocal communication of emotion: A review of research paradigms. *Speech Communication*. 2003, vol. 40, iss. 1–2, pp. 227–256, ISSN 0167-6393.

[11] DARWIN, Ch. The expression of the emotions in man and animals. Oxford University Press, 1998.

[12] PLUTCHIK, R. Emotion: A Psychoevolutionary Synthesis. Harper and Row, New York, 1980.

[13] VERVERIDIS, D. and C. KOTROPOULOS. Emotional speech recognition: Resources, features, and methods. *Speech Communication*. 2006, vol. 48, no. 9, pp. 1162–1181. DOI: 10.1016/j.specom.2006.04.003. ISSN 0167-6393.

[14] ANSCOMBE, E. and P. T. GEACH. *Descartes Philosophical Writings.* 1970, 2nd. ed., Melbourne.

[15] COWIE, R. and R. R. CORNELIUS. Describing the emotional states that are expressed in speech. *Speech Communication*. 2003, vol. 40, no. 1, pp. 5–32. DOI: 10.1016/S0167-6393(02)00071-7. ISSN 01676393.

[16] ECKMAN, P. An argument for basic emotions. *Cognition Emotion*. 1992, vol. 6, iss. 3, pp. 169–200.

[17] Buck, R. The biological affects: a typology. *Psychological review*. 1999, vol. 106, no.2, pp. 301–336.

[18] NAKATSU, R, J. NICHOLSON and N. TOSA. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-Based Systems*. 2000, vol. 13, iss. (7–8), pp. 497–504. DOI: 10.1016/S0950-7051(00)00070-8. ISSN 0950-7051.

[19] Rahurkar, M., J. H. L. Hansen. Frequency band analysis for stress detection using a Teager energy operator based feature. In: *Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02)*. 2002, vol. 3, pp. 2021–2024.

[20] PICARD, R. W., E. VYZAS and J. HEALEY. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 23, pp. 10, pp. 1175–1191. DOI: 10.1109/34.954607. ISSN 0162–8828.

[21] WAGNER, J., J. KIM and E. ANDRE. From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In: *IEEE International Conference on Multimedia and Expo*. Amsterdam: IEEE, 2005, pp. 940-943. DOI: 10.1109/ICME.2005.1521579. ISBN 0-7803-9331-7.

[22] SLANEY, M. and G. MCROBERTS. BabyEars: A recognition system for affective vocalizations. *Speech Communication*. 2003, vol. 39, iss. 3-4, pp. 367–384. DOI: 10.1016/S0167-6393(02)00049-3. ISSN 0167-6393.

[23] FRANCE, D. J., R. G. SHIAVI, S. SILVERMAN, M. SILVERMAN and M. WILKES. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*. vol. 47, iss. 7, pp. 829–837. DOI: 10.1109/10.846676. ISSN 0018-9294.

[24] LEE, C. M. and S. S. NARAYANAN. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*. 2005, vol. 13, iss. 2, pp. 293–303. DOI: 10.1109/TSA.2004.838534. ISSN 1063–6676.

[25] NWE, T. L., S. W. FOO and L. C. DE SILVA. Speech emotion recognition using hidden Markov models. *Speech Communication*. 2003, vol. 41, iss. 4, pp. 603–623. DOI: 10.1016/S0167-6393(03)00099-2. ISSN 0167-6393.

[26] BREAZEAL, C. and L. ARYANANDA. Recognition of Affective Communicative Intent in Robot-Directed Speech. *Autonomous Robots*. 2002, vol. 12, iss. 1, pp. 83–104. DOI: 10.1023/A:1013215010749. ISSN 0929-5593.

[27] DELLER, J. R., J. H. L. HANSEN and J. G. PROAKIS. *Discrete-time processing of speech signals*. 1st ed. New York: Wiley-IEEE Press, 2000. ISBN 07-803-5386-2.

[28] AYADI, E. M., M. S. KAMEL and F. KARRAY. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*. 2011, vol. 44, iss. 3, pp. 572–587. DOI: 10.1016/j.patcog.2010.09.020. ISSN 00313203.

[29] SCHULLER, B. and G. RIGOLL. Recognising interest in conversational speech - comparing bag of frames and supra-segmental features. In: *INTERSPEECH 2009*. Brighton: ISCA, 2009, pp. 1999–2000. ISBN 978-1-61567-692-7.

[30] NWE, T. L., S. W. FOO and L. C. DE SILVA. Classification of stress in speech using linear and nonlinear features. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Hong Kong: IEEE, 2003, pp. 9–12. DOI: 10.1109/ICASSP.2003.1202281. ISBN 0-7803-7663-3.

[31] SCHULLER, B., G. RIGOLL and M. LANG. Hidden Markov model-based speech emotion recognition. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03)*. Hong Kong: IEEE, 2003, pp. 1–4. DOI: 10.1109/ICASSP.2003.1202279. ISBN 0-7803-7663-3.

[32] LEE, C. M., S. YILDIRIM, M. BULUT, A. KAZEMZADEH, C. BUSSO, Z. DENG, S. LEE and S. S. NARAYANAN. Emotion recognition based on phoneme classes. In: *Proceedings of international conference spoken language processing*. Jeju Island: ICSLP, 2004, pp. 889–892.

[33] WANG, Y., S. DU and Y. ZHAN. Adaptive and Optimal Classification of Speech Emotion Recognition. In: *Fourth International Conference on Natural Computation*. Jinan: IEEE, 2008, pp. 407–411. DOI: 10.1109/ICNC.2008.713. ISBN 978-0-7695-3304-9.

[34] SCHULLER, B., A. BATLINER, D. SEPPI, S. STEIDL, T. VOGT, J. WAGNER, L. DEVILLERS, L. VIDRASCU, N. AMIR and L. KESSOUS, V. AHARONSON. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: *Proceedings of INTERSPEECH*. Antwerp: ISCA, 2007, pp. 2253–2256, ISSN 1990-9772.

[35] LUGGER, M., and B. YANG. The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Honolulu: IEEE, 2007, pp. 17–20. DOI: 10.1109/ICASSP.2007.367152. ISBN 1-4244-0727-3.

[36] FIROZ, A. S., V. R. KRISHMAN, R. A. SUKUMAR, A. JAYAKUMAR and B. P. ANTO. Speaker Independent Automatic Emotion Recognition from Speech: A Comparison of MFCCs and Discrete Wavelet Transforms. In: *International Conference on Advances in Recent Technologies in Communication and Computing*. Kottayam: IEEE, 2009, pp. 528–531. DOI: 10.1109/ARTCom.2009.231. ISBN 978-1-4244-5104-3.

[37] WU, CH.-H. and W.-B. LIANG. Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels. *IEEE Transactions on Affective Computing*. 2011, vol. 2, iss. 1, pp. 10–21. DOI: 10.1109/T-AFFC.2010.16. ISSN 1949-3045.

[38] *Encyclopedia of machine learning*. London: Springer, 2010, pp. 257–258. ISBN 9780387301648.

[39] ANAGNOSTOPOULOS, Ch.-N., T. ILIOU and I. GIANNOUKOS. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*. 2015, vol. 43, no. 2, pp. 155–177. DOI: 10.1007/s10462-012-9368-5. ISSN 0269-2821.

[40] ZHOU, Y., Y. SUN, L. YANG and Y. YAN. Applying Articulatory Features to Speech Emotion Recognition. In: *International Conference on Research Challenges in Computer Science*. Shanghai: IEEE, 2009, pp. 73–76. DOI: 10.1109/ICRCCS.2009.26. ISBN 978-1-4244-5409-9.

[41] WANG, S., X. LING, F. ZHANG and J. TONG. Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network. In: *International Conference on Measuring Technology and Mechatronics Automation*. IEEE, 2010, pp. 437–440. DOI: 10.1109/ICMTMA.2010.523. ISBN 978-1-4244-5001-5.

[42] CHENG, X. M., P. Y. CHENG and L. ZHAO. A Study on Emotional Feature Analysis and Recognition in Speech Signal. In: *International Conference on Measuring Technology and Mechatronics Automation*. IEEE, 2009, pp. 418–420. DOI: 10.1109/ICMTMA.2009.89. ISBN 978-0-7695-3583-8.

[43] SCHULLER, B., B. VLASENKO, F. EYBEN, M. WOLLMER, A. STUHLSATZ, A. WENDEMUTH and G. RIGOLL. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*. 2010, vol. 1, no. 2, pp. 119–131. DOI: 10.1109/T-AFFC.2010.8. ISSN 1949-3045.

[44] YANG, C., L. JI and G. LIU. Study to Speech Emotion Recognition Based on TWINsSVM. In: *Fifth International Conference on Natural Computation*. IEEE, 2009, pp. 312–316. DOI: 10.1109/ICNC.2009.464. ISBN 978-0-7695-3736-8.

[45] ATASSI, H. and A. ESPOSITO. A Speaker Independent Approach to the Classification of Emotional Vocal Expressions. In: *20th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*. Washington: IEEE, 2008, pp. 147–152. DOI: 10.1109/ICTAI.2008.158. ISBN 978-0-7695-3440-4.

[46] YUN, S. and C. D. YOO. Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei: IEEE, 2009, pp. 4169–4172. DOI: 10.1109/ICASSP.2009.4960547. ISBN 978-1-4244-2353-8.

[47] NOGUEIRAS, A., et al. Speech emotion recognition using hidden Markov models. In: *7th European Conference on Speech Communication and Technology*. Aalborg: ISCA, 2001.

[48] CEN, L., W. SER and Z. L. YU, Speech Emotion Recognition Using Canonical Correlation Analysis and Probabilistic Neural Network. *Seventh International Conference on Machine Learning and Applications*. San Diego: IEEE, 2008, pp. 859–862. DOI: 10.1109/ICMLA.2008.85. ISBN 978-0-7695-3495-4.

[49] ILIOU, T. and C.-N. ANAGNOSTOPOULOS. Comparison of Different Classifiers for Emotion Recognition. In: *13th Panhellenic Conference on Informatics*. Corfu: IEEE, 2009, pp. 102–106. DOI: 10.1109/PCI.2009.7. ISBN 978-0-7695-3788-7.

[50] HAN, K., D. YU and I. TASHEV. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In: *INTERSPEECH*. Singapure: ISCA, 2014, pp. 223–227.

[51] ERDAL, M., M. KACHELE and F. SCHWENKER. Emotion Recognition in Speech with Deep Learning Architectures. In: *7th IAPR TC3 Workshop*. Ulm: Springer International Publishing, 2016, pp. 298–311. DOI: 10.1007/978-3-319-46182-3_25. ISBN 978-3-319-46181-6.

[52] LI, L., Y. ZHAO, D. JIANG, Y. ZHANG, F. WANG, I. GONZALEZ, E. VALENTIN and H. SAHLI. Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In: *Humaine Association Conference on Affective Computing and Intelligent Interaction*. Geneva: IEEE, 2013, pp. .312–317. DOI: 10.1109/ACII.2013.58. ISBN 978-0-7695-5048-0.

[53] WERNER, S. and E. KELLER. *Fundamentals of speech synthesis and speech recognition: basic concepts, state of the art, and future challenges*. New York: Wiley, 1994, pp. 23–40. ISBN 0-471-94449-1.

[54] DUTOIT, T. *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Springer Netherlands, 1997. ISBN 978-940-1157-308.

[55] QUATIERI, T. F. *Discrete-time speech signal processing principles and practice.* Upper Saddle River, N.J: Prentice Hall, 2002. ISBN 978-013-2442-138.

[56] LUGGER, M. and B. YANG. *Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features.* Speech recognition: technologies and applications. 1st. ed. Vienna: I-Tech Education and Publishing, 2008, pp. 395–410. ISBN 9789537619299.

[57] BOERSMA, P. Praat, a system for doing phonetics by computer. *Glot international.* 2002, vol. 5, no. 9/10, pp. 341–345. ISSN 1381-3439.

[58] MARKEL, J. D. and H. G. AUGUSTINE. *Linear prediction of speech.* Berlin: Springer, 1976. ISBN 978-364-2662-881.

[59] MARYN, Y., N. ROY, M. DE BODT, P. VAN CAUWENBERGE and P. CORTHALS. Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America.* 2009, vol. 126, no. 5, pp. 2619–2634. DOI: 10.1121/1.3224706. ISSN 00014966.

[60] FRAILE, R. and Juan I. GODINO-LLORENTE. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control.* 2014, vol. 14, pp. 42–54. DOI: 10.1016/j.bspc.2014.07.001. ISSN 17468094.

[61] KIM, H. K., K. C. KIM and H.S. LEE. Enhanced distance measure for LSP-based speech recognition. *Electronics Letters.* 1993, vol. 29, no. 16, 1463-1465. DOI: 10.1049/el:19930979. ISSN 00135194.

[62] JUNQUA, J.-C., and J.-P. Haton. *Robustness in Automatic Speech Recognition Fundamentals and Applications.* Boston: Springer US, 1996. ISBN 978-146-1312-970

[63] SWIETOJANSKI, P., A. GHOSHAL and S. RENALS, Convolutional Neural Networks for Distant Speech Recognition. *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014. DOI: 10.1109/LSP.2014.2325781

[64] EYBEN, F, M WOELLMER a B SCHULLER. *OpenSMILE: the Munich open Speech and Music Interpretation by Large Space Extraction toolkit.* 2010 Institute for Human-Machine Communication Technische Universitaet Muenchen (TUM) D-80333 Munich, Germany http://www.mmk.ei.tum.de

[65] JEON, J. H., R. XIA and Y. LIU. Level of interest sensing in spoken dialog using decision-level fusion of acoustic and lexical evidence. Computer Speech. 2014, 28(2), 420-433. DOI: 10.1016/j.csl.2013.09.005. ISSN 08852308.

[66] JOLLIFFE, I. T. *Principal component analysis.* 2nd ed. New York: Springer, 2002. ISBN 03-879-5442-2.

[67] ABDI, H. and L. J. WILLIAMS. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2010, vol. 2, no. 4, pp. 433–459. DOI: 10.1002/wics.101. ISSN 19395108.

[68] MIRANDA, A. A., Y.-A. LE BORGNE and G. BONTEMPI. New Routes from Minimal Approximation Error to Principal Components. *Neural Processing Letters*. 2008, vol. 27, no. 3, pp. 197–207. DOI: 10.1007/s11063-007-9069-2. ISSN 1370-4621.

[69] MARQUARDT, D. W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*. 1963, vol. 11, no. 2, pp. 431–441. DOI: 10.1137/0111030. ISSN 0368-4245.

[70] NEUBRGER, T. and A. BEKE. Automatic Laughter Detection in Spontaneous Speech Using GMM–SVM Method. *Lecture Notes in Computer Science*. 2013. pp. 113. DOI: 10.1007/978-3-642-40585-3-15.

[71] KRAJEWSKI, J., SCHNIEDER, S., SOMMER, D., BATLINER, A. and SCHULLER, B. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*. 2012, vol. 84, pp. 65–75. DOI: 10.1016/j.neucom.2011.12.021.

[72] CORTES, C. and V. VAPNIK. Support-Vector Networks. *Machine Learning*. 1995, vol. 20, iss. 3, pp. 273–297. DOI: 10.1023/A:1022627411411. ISSN 08856125.

[73] SCHOLKOPF, B. and A. J. SMOLA. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, Mass.: MIT Press, 2002. Adaptive computation and machine learning. ISBN 0-262-19475-9

[74] PROVOST, F. and R. KOHAVI. Guest Editors' Introduction: On Applied Research in Machine Learning. *Machine Learning*. 1998, vol. 30, iss. 2, pp. 127–132. DOI: 10.1023/A:1007442505281. ISSN 0885-6125.

[75] Powers, D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*. 2011, vol. 2, no. 1, pp. 37–63.

[76] GRZESIAK, W. and D. ZABORSKI. Examples of the Use of Data Mining Methods in Animal Breeding. *Data Mining Applications in Engineering and Medicine*. 2012. DOI: 10.5772/50893. ISBN 978-953-51-0720-0.

[77] Berlin Database of Emotional Speech [online]. [cit. 2016-11-03]. Available at: http://emodb.bilderbar.info/start.html

[78] GOVINDARAJU, V. and C. R. RAO. *Machine learning: theory and applications*. Boston: Elsevier/North Holland, 2013. Handbook of statistics (Amsterdam, Netherlands), vol. 31. ISBN 04-445-3859-3.

[79] Ruta, D. and Gabrys, B., An Overview of Classifier Fusion Methods. *Computing and Information Systems*. 2000, vol. 7, no. 1, pp. 1-10. ISSN 1352-9404.

[80] WOZNIAK, M., M. GRANA and E. CORCHADO. A survey of multiple classifier systems as hybrid systems. *Information Fusion*. 2014, vol. 16, pp. 3–17. DOI: 10.1016/j.inffus.2013.04.006. ISSN 15662535.

[81] MA, A. J., P. C. YUEN and J.-H. LAI. Linear Dependency Modeling for Classifier Fusion and Feature Combination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013, vol. 35, iss. 5, pp. 1135–1148. DOI: 10.1109/TPAMI.2012.198. ISSN 0162-8828

[82] PAPPAS, D., I. ANDROUTSOPOULOS and H. PAPAGEORGIOU. Anger detection in call center dialogues. In: *6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. Gyor: IEEE, 2015, pp. 139–144. DOI: 10.1109/CogInfoCom.2015.7390579. ISBN 978-1-4673-8129-1.

[83] VAUDABLE, C. and L. DEVILLERS. Negative emotions detection as an indicator of dialogs quality in call centers. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto: IEEE, 2012, pp. 5109–5112. DOI: 10.1109/ICASSP.2012.6289070. ISBN 978-1-4673-0046-9.

[84] ATASSI, H., Z SMEKAL and A. ESPOSITO. Emotion recognition from spontaneous Slavic speech. In: *3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. Kosice: IEEE, 2012. pp.389–394 DOI: 10.1109/CogInfoCom.2012.6422011. ISBN 978-1-4673-5187-4.

CANDIDATE'S RESEARCH CITED IN THIS WORK

[Par12] PARTILA, P., M. VOZNAK, M. MIKULEC and J. ZDRALEK. Fundamental frequency extraction method using central clipping and its importance for the classification of emotional state.*Advances in electrical and electronic engineering.* 2012, vol. 10, no. 4, pp. 270–275. ISSN 1804-3119. SCOPUS, SJR 0.164 (2012/Q4)

[Par15] PARTILA, P., M. VOZNAK and J. TOVAREK. Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System. *The Scientific World Journal.* 2015, pp. 1–7. DOI: 10.1155/2015/573068. ISSN 2356-6140. SCOPUS, SJR 0.315 (2015/Q2).

[Par16] PARTILA, P., J. TOVAREK and M. VOZNAK. Self-organizing map classifier for stressed speech recognition. In: *Machine Intelligence and Bio-Inspired Computation: Theory and Applications X.* vol. 9850. Baltimore, USA: SPIE, 2016, art. no. 98500A. DOI: 10.1117/12.2224253. ISBN 978-151060091-1. WoS, SCOPUS, SJR 0.216 (2015).

[Par14] PARTILA, P., J. TOVAREK, J. FRNDA, M. VOZNAK, M. PENHAKER and T. PETEREK. Emotional Impact on Neurological Characteristics and Human Speech. In: *1st Euro-China Conference on Intelligent Data Analysis and Applications.* Shenzhen, China: Springer, 2014, pp. 527–533. DOI: 10.1007/978-3-319-07773-4_52. ISBN 978-331907772-7. SCOPUS, SJR 0.149 (2014/Q4).

[Uhr16] UHRIN, D., Z. Chmelikova, J. Tovarek, P. Partila, M. Voznak. One approach to design of speech emotion database. In: *Proceedings of SPIE Vol. 9850. Machine Intelligence and Bio-inspired Computation: Theory and Applications X.* Baltimore, USA: SPIE, 2016. DOI. 10.1117/12.2227067. WoS, SCOPUS, SJR 0.216 (2015)

[Tov16] TOVAREK, J., P. PARTILA, J. ROZHON, M. VOZNAK, J. SKAPA, D. UHRIN and Z. CHMELIKOVA. Optimization of multilayer neural network parameters for speaker recognition. In: *Machine Intelligence and Bio-Inspired Computation: Theory and Applications X.* Baltimore, USA: SPIE, 2016, art. no. 98500. DOI: 10.1117/12.2223545. ISBN 978-151060091-1. WoS, SCOPUS, SJR 0.216 (2015)

[Uhr14] UHRIN, D., P. PARTILA, M. VOZNAK, Z. CHMELIKOVA, M. HLOZAK and L. ORCIK. Design and implementation of Czech database of speech emotions. In: *22nd Telecommunications Forum Telfor (TELFOR).* Belgrade, Republic of Serbia: IEEE, 2014, pp. 529–532. DOI: 10.1109/TELFOR.2014.7034463. ISBN 978-1-4799-6191-7. SCOPUS, IEEE-Xplore

[Tov15] TOVAREK, J., P. PARTILA, M. VOZNAK, M. MIKULEC and M. MEHIC. Detection of cardiac activity changes from human speech. In: *Independent Component Analyses, Compressive Sampling, Large Data Analyses (LDA), Neural Networks, Biosystems, and Nanoengineering XIII.* Baltimore, USA: SPIE, 2015, art. no. 94960V. DOI: 10.1117/12.2177282. ISBN 978-162841612-1. WoS, SCOPUS, SSJR 0.216 (2015)

[Par14b] PARTILA, P., J. TOVAREK, M VOZNAK and J. SAFARIK. Classification Methods Accuracy for Speech Emotion Recognition System. In: *Advances in Intelligent Systems and Computing*. Ostrava, Czech Republic: Springer, 2014, pp. 439–447. DOI: 10.1007/978-3-319-07401-6_44. ISBN 978-331907400-9. WoS, SCOPUS, SJR 0.149 (2014/Q4)

[Par14c] PARTILA, P., M. VOZNAK, T. PETEREK, M. PENHAKER, V. NOVAK, J. TOVAREK, M. MEHIC and L. VOJTECH. Impact of human emotions on physiological characteristics. In: *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XII*. Baltimore, USA: SPIE, 2014, art. no. 91180W. DOI: 10.1117/12.2050679. ISBN 978-162841055-6. WoS, SCOPUS, SJR 0.22 (2014)

[Voz13b] VOZNAK, M., P. PARTILA, M. MEHIC and S. JAKOVLEV. Recognizing emotions from human speech using 2-D neural classifier and influence the selection of input parameters on its accuracy. In: *21st Telecommunications Forum Telfor (TELFOR)*. Belgrade, Republic of Serbia: IEEE, 2013, pp. 482–485. DOI: 10.1109/TELFOR.2013.6716272. ISBN 978-1-4799-1420-3. WoS, SCOPUS, IEEE-Xplore.

[Par13] PARTILA, P. and M. VOZNAK. Speech Emotions Recognition Using 2-D Neural Classifier. In: Advances in Intelligent Systems and Computing. Ostrava, Czech Republic: Springer, 2013, pp. 221–231. DOI: 10.1007/978-3-319-00542-3_23. ISBN 978-331900541-6. SCOPUS, SJR 0.145 (2013/Q4)

LIST OF CANDIDATE'S RESEARCH RESULTS AND ACTIVITIES

Publication activities

I provide the following list indexed results in relevant scientific databases, in order to document my research activities within the entire period of my doctoral study:

- records in **Elsevier Scopus**: 30 (24 conference papers, 6 articles in journals)
- records in **ISI Web of Knowledge**: 18 (18 conference papers)
- records in **IEEE-Xplore**: 5
- h-index according to ISI/WoS: 1 (2 citations)
- h-index according to Scopus: 3 (30 citations)

Project memberships and participations

- member of INDECT team, **Intelligent information system supporting observation, searching and detection for security of citizens in urban environment**. The 7FP EU (2009-2014) under Grant Agreement No. 218086, conducted by AGH Cracow.
- member of Research Programme No. 5, **IT4Innovations**, Supercomputing centre, Czech National Centre of Excellence, Ostrava, 2011-2015
- member of research team, **Security of mobile devices and communication**, Technology Agency of the Czech Republic, under grant TF01000091 (2015-2017).
- member of research team, **Development of human resources in research and development of latest soft computing methods and their application in practice project**, under grant CZ.1.07/2.3.00/20.0072, funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic (2012-2014)
- Specific research, SGS FEI VSB-TU Ostrava, project SP2016/170, **Knowledge retrieval in communications networks, modelling and simulation - II**.
- Specific research, SGS FEI VSB-TU Ostrava, project SP2015/82, **Knowledge retrieval in communications networks, modelling and simulation - I**.
- Specific research, SGS FEI VSB-TU Ostrava, project SP2014/72, **Research on the impact of atmospheric phenomenas on the transmission in radio channels**.
- member of team in development project, FEI VSB-TU Ostrava, project FRVS2013/1467, **Creating a new laboratory tasks in the integration of voice with enterprise of information systems**.
- Specific research, SGS FEI VSB-TU Ostrava, project SP2013/94, **Research on the impact of the environment on the properties of the radio channel and the**

development of new approaches to the evaluation of the quality of service (QoS) in 4G multimedia networks**.

- Specific research, SGS FEI VSB-TU Ostrava, project SP2012/180, **Changes in conditions of radio signals propagation due to the weather**.

Intership

Erasmus scholarship - **Ankara University, Faculty of Engineering, Electrical and Electronics Engineering Department**, Ankara (Turkey), Study in field of the signal processing and speech processing. Intership was focused on the design of a system for speech emotion recognition in Winter semestr of ac. year 2013–2014.

Results directly related to the topic of dissertation indexed on Web of Science or in Elsevier Scopus (13 publications)

- PARTILA, P., M. VOZNAK, M. MIKULEC and J. ZDRALEK. Fundamental frequency extraction method using central clipping and its importance for the classification of emotional state.*Advances in electrical and electronic engineering*. 2012, vol. 10, no. 4, pp. 270–275. ISSN 1804-3119. SCOPUS, SJR 0.164 (2012/Q4)
- PARTILA, P., M. VOZNAK and J. TOVAREK. Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System. *The Scientific World Journal*. 2015, pp. 1–7. DOI: 10.1155/2015/573068. ISSN 2356-6140. SCOPUS, SJR 0.315 (2015/Q2).
- PARTILA, P., J. TOVAREK and M. VOZNAK. Self-organizing map classifier for stressed speech recognition. In: *Machine Intelligence and Bio-Inspired Computation: Theory and Applications X*. vol. 9850. Baltimore, USA: SPIE, 2016, art. no. 98500A. DOI: 10.1117/12.2224253. ISBN 978-151060091-1. WoS, SCOPUS, SJR 0.216 (2015).
- PARTILA, P., J. TOVAREK, J. FRNDA, M. VOZNAK, M. PENHAKER and T. PETEREK. Emotional Impact on Neurological Characteristics and Human Speech. In: *1st Euro-China Conference on Intelligent Data Analysis and Applications*. Shenzhen, China: Springer, 2014, pp. 527–533. DOI: 10.1007/978-3-319-07773-4_52. ISBN 978-331907772-7. SCOPUS, SJR 0.149 (2014/Q4).
- UHRIN, D., Z. Chmelikova, J. Tovarek, P. Partila, M. Voznak. One approach to design of speech emotion database. In: *Proceedings of SPIE Vol. 9850. Machine Intelligence and Bio-inspired Computation: Theory and Applications X*. Baltimore, USA: SPIE, 2016. DOI. 10.1117/12.2227067. WoS, SCOPUS, SJR 0.216 (2015)
- TOVAREK, J., P. PARTILA, J. ROZHON, M. VOZNAK, J. SKAPA, D. UHRIN and Z. CHMELIKOVA. Optimization of multilayer neural network parameters for speaker recognition. In: *Machine Intelligence and Bio-Inspired Computation: Theory*

and Applications X. Baltimore, USA: SPIE, 2016, art. no. 98500. DOI: 10.1117/12.2223545. ISBN 978-151060091-1. WoS, SCOPUS, SJR 0.216 (2015)
- UHRIN, D., P. PARTILA, M. VOZNAK, Z. CHMELIKOVA, M. HLOZAK and L. ORCIK. Design and implementation of Czech database of speech emotions. In: *22nd Telecommunications Forum Telfor (TELFOR)*. Belgrade, Republic of Serbia: IEEE, 2014, pp. 529–532. DOI: 10.1109/TELFOR.2014.7034463. ISBN 978-1-4799-6191-7. SCOPUS, IEEE-Xplore
- TOVAREK, J., P. PARTILA, M. VOZNAK, M. MIKULEC and M. MEHIC. Detection of cardiac activity changes from human speech. In: *Independent Component Analyses, Compressive Sampling, Large Data Analyses (LDA), Neural Networks, Biosystems, and Nanoengineering XIII*. Baltimore, USA: SPIE, 2015, art. no. 94960V. DOI: 10.1117/12.2177282. ISBN 978-162841612-1. WoS, SCOPUS, SSJR 0.216 (2015)
- PARTILA, P., J. TOVAREK, M VOZNAK and J. SAFARIK. Classification Methods Accuracy for Speech Emotion Recognition System. In: *Advances in Intelligent Systems and Computing*. Ostrava, Czech Republic: Springer, 2014, pp. 439–447. DOI: 10.1007/978-3-319-07401-6_44. ISBN 978-331907400-9. WoS, SCOPUS, SJR 0.149 (2014/Q4)
- PARTILA, P., M. VOZNAK, T. PETEREK, M. PENHAKER, V. NOVAK, J. TOVAREK, M. MEHIC and L. VOJTECH. Impact of human emotions on physiological characteristics. In: *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XII*. Baltimore, USA: SPIE, 2014, art. no. 91180W. DOI: 10.1117/12.2050679. ISBN 978-162841055-6. WoS, SCOPUS, SJR 0.22 (2014)
- VOZNAK, M., P. PARTILA, M. MEHIC and S. JAKOVLEV. Recognizing emotions from human speech using 2-D neural classifier and influence the selection of input parameters on its accuracy. In: *21st Telecommunications Forum Telfor (TELFOR)*. Belgrade, Republic of Serbia: IEEE, 2013, pp. 482–485. DOI: 10.1109/TELFOR.2013.6716272. ISBN 978-1-4799-1420-3. WoS, SCOPUS, IEEE-Xplore.
- PARTILA, P. and M. VOZNAK. Speech Emotions Recognition Using 2-D Neural Classifier. In: Advances in Intelligent Systems and Computing. Ostrava, Czech Republic: Springer, 2013, pp. 221–231. DOI: 10.1007/978-3-319-00542-3_23. ISBN 978-331900541-6. SCOPUS, SJR 0.145 (2013/Q4)
- VOZNAK, M., P. PARTILA, M. PENHAKER, T. PETEREK, K. TOMALA, F. REZAC and J. SAFARIK. Emotional state and its impact on voice authentication accuracy. In: *Conference on Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI*. Baltimore, USA: SPIE, 2013, art. no. 875006. DOI: 10.1117/12.2015719. ISBN 978-081949541-9

Other results indexed in Elsevier Scopus

I achieved next results not directly related to the dissertation within my study period and which are indexed in Elsevier.

- MEHIC, M., P. FAZIO, M. VOZNAK, P. PARTILA, D. KOMOSNY, J. TOVAREK and Z. CHMELIKOVA. On using multiple routing metrics with destination sequenced distance vector protocol for MultiHop wireless ad hoc networks. In: *Modeling and Simulation for Defense Systems and Applications XI*. Baltimore, USA: SPIE, 2016, art. no. 98480F. DOI: 10.1117/12.2223671. ISBN 978-151060089-8.
- MIKULEC, M., M. VOZNAK, M. FAJKUS, P. PARTILA, J. TOVAREK and Z. CHMELIKOVA. Building GSM network in extreme conditions. In: *Modeling and Simulation for Defense Systems and Applications X*. Baltimore, USA: SPIE, 2015, art. no. 94780K. DOI: 10.1117/12.2177027. ISBN 978-162841594-0.
- MEHIC, M., P. PARTILA, J. TOVAREK and M. VOZNAK. Calculation of key reduction for B92 QKD protocol. In: *Quantum Information and Computation XIII*. Baltimore, USA: SPIE, 2015, art. no. 95001J. DOI: 10.1117/12.2177149. ISBN 978-162841616-9.
- MEHIC, M., M. VOZNAK, J. SAFARIK, P. PARTILA and M. MIKULEC. Using DNS amplification DDoS attack for hiding data. In: *Mobile Multimedia/Image Processing, Security, and Applications*. Baltimore, USA: SPIE, 2014, art. no. 91200R. DOI: 10.1117/12.2050700. ISBN 978-162841057-0.
- MIKULEC, M., M. VOZNAK, J. SAFARIK, P. PARTILA, J. ROZHON and M. MEHIC. Interactive video audio system: Communication server for INDECT portal. In: *Mobile Multimedia/Image Processing, Security, and Applications*. Baltimore, USA: SPIE, 2014, art. no. 91200U. DOI: 10.1117/12.2050690. ISBN 978-162841057-0.
- VOZNAK, M., J. ROZHON, P. PARTILA, J. SAFARIK, M. MIKULEC and M. MEHIC. Predictive model for determining the quality of a call. In: *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XII*. Baltimore, USA: SPIE, 2014, art. no. 91180Y. DOI: 10.1117/12.2050661. ISBN 978-162841055-6.
- SAFARIK, J., M. VOZNAK, M. MEHIC, P. PARTILA and M. MIKULEC. Neural network classifier of attacks in IP telephony. In: *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XII*. Baltimore, USA: SPIE, 2014, art. no. 91180X. DOI: 10.1117/12.2050671. ISBN 978-162841055-6.
- REZAC, F., M. VOZNAK, P. PARTILA and K. TOMALA. Interactive video audio system and its performance evaluation. In: *36th International Conference on Telecommunications and Signal Processing (TSP)*. Rome, Italy: IEEE, 2013, pp. 43–46. DOI: 10.1109/TSP.2013.6613888. ISBN 978-1-4799-0404-4.

- TOMALA, K., J. ROZHON, F. REZAC, P. PARTILA, M. VOZNAK and M. MIKULEC. Interactive multimedia module into danger alert communication system. In: *36th International Conference on Telecommunications and Signal Processing (TSP)*. Rome, Italy: IEEE, 2013, pp. 195–198. DOI: 10.1109/TSP.2013.6613918. ISBN 978-1-4799-0404-4.
- MIKULEC, M., L. KAPICAK, M. VOZNAK, P. PARTILA, K. TOMALA, P. NEVLUD and J. ZDRALEK. Implementation of voice, SMS and MMS services into Interactive Video Audio System (IVAS). In: *36th International Conference on Telecommunications and Signal Processing (TSP)*. Rome: IEEE, 2013, s. 204-207. DOI: 10.1109/TSP.2013.6613920. ISBN 978-1-4799-0404-4.
- REZAC, F., M. VOZNAK, J. SAFARIK, P. PARTILA and K. TOMALA. Security solution against denial of service attacks in BESIP system. In: *Mobile Multimedia/Image Processing, Security, and Applications*. Baltimore, USA: SPIE, 2013, art. no. 87550Y. DOI: 10.1117/12.2015379. ISBN 978-081949546-4.
- SAFARIK, J., M. VOZNAK, F. REZAC, P. PARTILA and K. TOMALA. Automatic analysis of attack data from distributed honeypot network. In: *Mobile Multimedia/Image Processing, Security, and Applications*. Baltimore, USA: SPIE, 2013, art. no. 875512. DOI: 10.1117/12.2015514. ISBN 978-081949546-4.
- REZAC, F., J. SAFARIK, M. VOZNAK, K. TOMALA and P. PARTILA. IP telephony based danger alert communication system and its implementation. In: *Mobile Multimedia/Image Processing, Security, and Applications*. Baltimore, USA: SPIE, 2013, art. no. 87550Z. DOI: 10.1117/12.2015423. ISBN 978-081949546-4.
- TOMALA, K., M. VOZNAK, P. PARTILA, F. REZAC and J. SAFARIK. Analysis and removing noise from speech using wavelet transform. In: *Conference on Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI*. Baltimore, USA: SPIE, 2013, art. no. 87500D. DOI: 10.1117/12.2015722. ISBN 978-081949541-9.
- PARTILA, P., M. VOZNAK, A. KOVAC and M. HALAS. Jitter buffer loss estimate for effective equipment impairment factor. *International Journal of Mathematics and Computers in Simulation*. 2013, vol. 7, no. 3, pp. 241–248. ISSN 1998-0159.
- SAFARIK, J., P. PARTILA, F. REZAC, L. MACURA and M. VOZNAK. Automatic Classification of Attacks on IP Telephony. *Advances in Electrical and Electronic Engineering*. 2013, vol. 11, iss. 6, pp. 481–486. DOI: 10.15598/aeee.v11i6.899. ISSN 1804-3119.
- PARTILA, P., M. KOHUT, M. VOZNAK, M. MIKULEC, J. SAFARIK and K. TOMALA. A Methodology for Measuring Voice Quality Using PESQ and Interactive Voice Response in the GSM Channel Designed by OpenBTS. *Advances in Electrical and Electronic Engineering*. 2013, vol. 11, iss. 5, pp. 380–386.
DOI: 10.15598/aeee.v11i5.894. ISSN 1804-3119.

A DATABSES - SUBJECTIVE EVALUATION ENVIRONMENT AND OVERVIEW LIST OF KNOWN DATABASES

Fig. A.1: Web environment for subjective evaluation of emoDBova database.

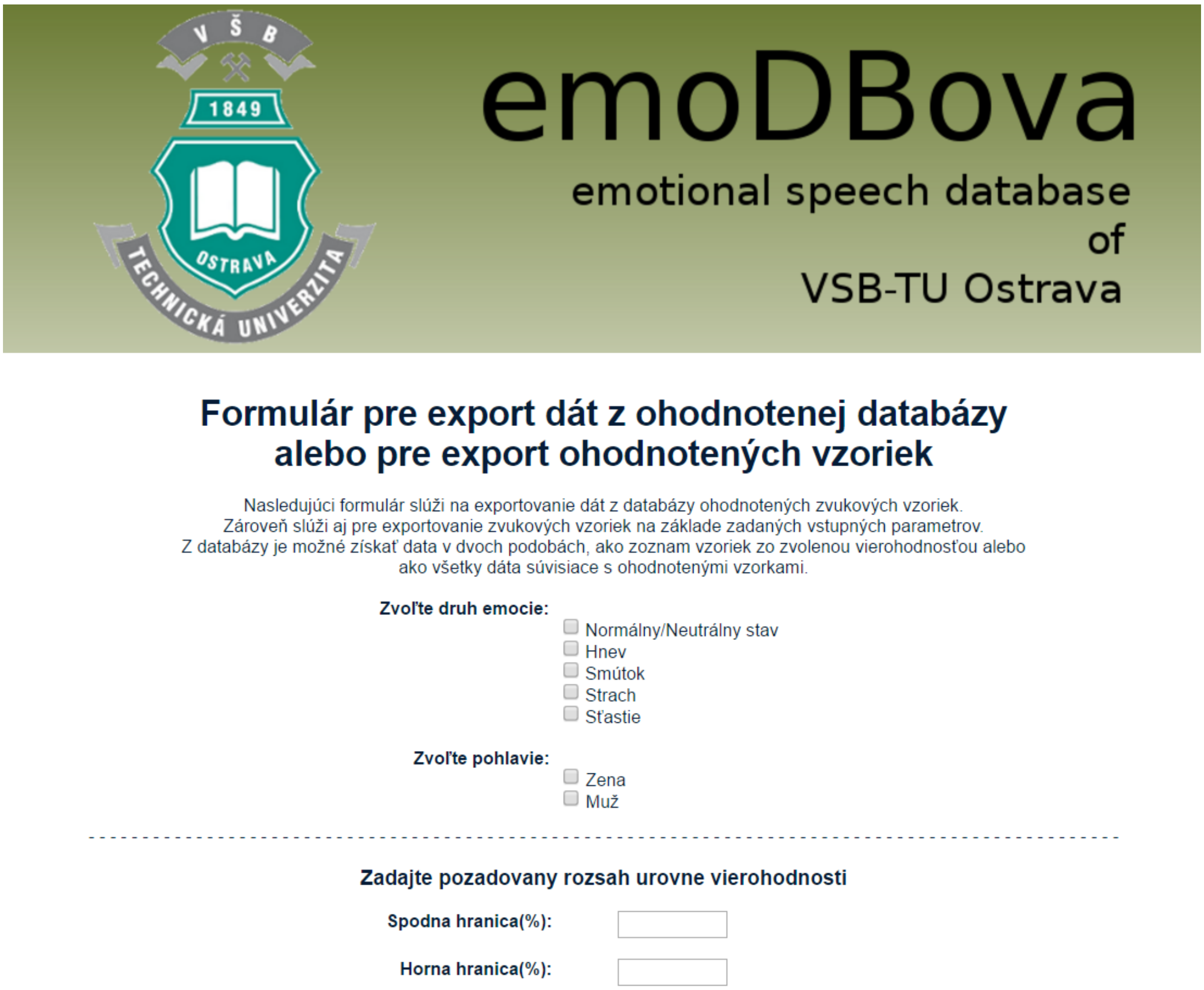


Fig. A.2: Extraction from emoDBova database.

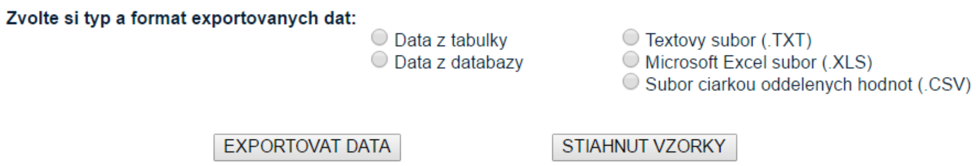


Fig. A.3: Extraction from emoDBova database - file formats.

```
        This is export of list_data in level of confidence from 0 to 100.

ref_id,value_of_veracity,level_of_veracity,gender_of_speaker,final_emotion
5_2_25_dav079,100.00,high,female,fear
1_1_12_PAV638,100.00,high,male,happiness
2_2_21_VAN434,100.00,high,female,anger
5_2_4_las129,100.00,high,female,anger
5_2_21_bed188,100.00,high,female,normal/neutral
1_1_1_Laz068,100.00,high,male,normal/neutral
3_1_14_PLA071,50.00,low,male,fear
4_1_4_JAW146,50.00,low,male,anger
4_1_3_krc082,100.00,high,male,happiness
4_1_23_PAV638,50.00,low,male,happiness
1_2_16_dav079,50.00,low,female,anger
1_1_8_HOL718,100.00,high,male,normal/neutral
2_1_7_bed188,100.00,high,male,anger
1_1_2_zbo062,50.00,low,male,fear
4_1_12_HAS113,50.00,low,male,fear
3_2_12_pob031,50.00,low,female,happiness
5_2_24_kou225,100.00,high,female,fear
4_1_2_las129,100.00,high,male,happiness
5_2_18_PLA071,50.00,low,female,anger
2_2_9_zbo062,50.00,low,female,anger
3_2_4_lak032,33.33,low,female,fear
3_2_4_sim285,100.00,high,female,sadness
3_1_24_CHR240,50.00,low,male,anger
5_2_24_zbo062,50.00,low,female,fear
2_2_10_zbo062,100.00,high,female,anger
5_2_25_han426,50.00,low,female,fear
1_1_3_han426,50.00,low,male,anger
4_1_16_POD204,50.00,low,male,happiness
```

Fig. A.4: Extraction from emoDBova database - records.

Tab. A.1: Listing of Emotional databases with additional information [13].

Reference	Language	Subjects	Other signals	Purpose	Emotions	Kind
Abelin and Allwood (2000)	Swedish	1 Native	-	Recognition	Ar, Fr, Jy, Sd, Se, Dt, Dom, Sy	Simulated
Alpert et al. (2001)	English	22 Patients, 19 healthy	-	Recognition	Dn, Nl	Natural
Alter et al. (2000)	German	1 Female	EEG	Recognition	Ar, Hs, Nl	Simulated
Ambrus (2000), Interface	English, Slovenian	8 Actors	LG	Synthesis	Ar, Dt, Fr, Nl, Se	Simulated
Amir et al. (2000)	Hebrew	40 Students	LG, M, G, H	Recognition	Ar, Dt, Fr, Jy, Sd	Natural
Ang et al. (2002)	English	Many	-	Recognition	An, At, Nl, Fd, Td	Natural
Banse and Scherer (1996)	German	12 Actors	V	Recognition	H/C Ar, Hs, Sd, ...	Simulated
Batliner et al. (2004)	German, English	51 Children	-	Recognition	Ar, Bm, Jy, Se	Elicited
Bulut et al. (2002)	English	1 Actress	-	Synthesis	Ar, Hs, Nl, Sd	Simulated
Burkhardt and Sendlmeier (2000)	German	10 Actors	V, LG	Synthesis	Ar, Fr, Jy, Nl, Sd, Bm, Dt	Simulated
Caldognetto et al. (2004)	Italian	1 Native	V, IR	Synthesis	Ar, Dt, Fr, Jy, Sd, Se	Simulated
Choukri (2003), Groningen	Dutch	238 Native	LG	Recognition	Unknown	Simulated
Chuang and Wu (2002)	Chinese	2 Actors	-	Recognition	Ar, Ay, Hs, Fr, Se, Sd	Simulated
Clavel et al. (2004)	English	18 From TV	-	Recognition	Nl, levels of Fr	Simulated
Cole (2005), Kids' Speech	English	780 Children	V	Recognition, Synthesis	Unknown	Natural
Cowie and Douglas-Cowie (1996), Belfast Structured	English	40 Native	-	Recognition	Ar, Fr, Hs, Nl, Sd	Natural
Douglas-Cowie et al. (2003), Belfast Natural	English	125 From TV	V	Recognition	Various	Semi-natural
Edgington (1997)	English	1 Actor	LG	Synthesis	Ar, Bm, Fr, Hs, Nl, Sd	Simulated
Engberg and Hansen (1996), DES	Danish	4 Actors	-	Synthesis	Ar, Hs, Nl, Sd, Se	Simulated
Fernandez and Picard (2003)	English	4 Drivers	-	Recognition	Nl, Ss	Natural

Fischer (1999), Verbmobil	German	58 Native	-	Recognition	Ar, Dn, Nl	Natural
France et al. (2000)	English	70 Patients, 40 healthy	-	Recognition	Dn, Nl	Natural
Gonzalez (1999)	English, Spanish	Unknown	-	Recognition	Dn, Nl	Elicited
Hansen (1996), SUSAS	English	32 Various	-	Recognition	Ar, Ld eff., Ss, Tl	Natural, Simulated
Hansen (1996), SUSC-0	English	18 Non-native	H, BP, R	Recognition	Nl, Ss	A-stress
Hansen (1996), SUSC-1	English	20 Native	-	Recognition	Nl, Ss	P-stress
Hansen (1996), DLP	English	15 Native	-	Recognition	Nl, Ss	C-stress
Hansen (1996), DCIEM	English	Unknown	-	Recognition	Nl, Sleep deprive	Elicited
Heuft et al. (1996)	German	3 Native	-	Synthesis	Ar, Fr, Jy, Sd, ...	Simulated, elicited
Iida et al. (2000), ESC	Japanese	2 Native	-	Synthesis	Ar, Jy, Sd	Simulated
Iriondo et al. (2000)	Spanish	8 Actors	-	Synthesis	Fr, Jy, Sd, Se, ...	Simulated
Kawanami et al. (2003)	Japanese	2 Actors	-	Synthesis	Ar, Hs, Nl, Sd	Simulated
Lee and Narayanan (2005)	English	Unknown	-	Recognition	Negative-positive	Natural
Liberman (2005), Emotional Prosody	English	Actors	-	Unknown	Antxy, H/C Ar, Hs, Nl, Pc, Sd, Se, ...	Simulated
Linnankoski et al. (2005)	English	13 Native	-	Recognition	An, Ar, Fr, Sd, ...	Elicited
Lloyd (1999)	English	1 Native	-	Recognition	Phonological stress	Simulated
Makarova and Petrushin (2002), RUSSLANA	Russian	61 Native	-	Recognition	Ar, Hs, Se, Sd, Fr, Nl	Simulated
Martins et al. (1998), BDFALA	Portuguese	10 Native	-	Recognition	Ar, Dt, Hs, Iy	Simulated
McMahon et al. (2003), ORESTEIA	English	29 Native	-	Recognition	Ae, Sk, Ss	Elicited
Montanari et al. (2004)	English	15 Children	V	Recognition	Unknown	Natural
Montero et al. (1999), SES	Spanish	1 Actor	-	Synthesis	Ar, Dt, Hs, Sd	Simulated
Mozziconacci and Hermes (1997)	Dutch	3 Native	-	Recognition	Ar, Bm, Fr, Jy, Iy, Nl, Sd	Simulated
Niimi et al. (2001)	Japanese	1 Male	-	Synthesis	Ar, Jy, Sd	Simulated
Nordstrand et al. (2004)	Swedish	1 Native	V, IR	Synthesis	Hs, Nl	Simulated
Nwe et al. (2003)	Chinese	12 Native	-	Recognition	Ar, Fr, Dt, Jy, ...	Simulated
Pereira (2000)	English	2 Actors	-	Recognition	H/C Ar, Hs, Nl, Sd	Simulated
Petrushin (1999)	English	30 Native	-	Recognition	Ar, Fr, Hs, Nl, Sd	Simulated , Natural

Polzin and Waibel (2000)	English	Unknown	-	Recognition	Ar, Fr, Nl, Sd	Simulated
Polzin and Waibel (1998)	English	5 Drama students	LG	Recognition	Ar, Fr, Hs, Nl, Sd	Simulated
Rahurkar and Hansen (2002), SOQ	English	6 Soldiers	H, R, BP, BL	Recognition	5 Stress levels	Natural
Scherer (2000b), Lost Luggage	Various	109 Passengers	V	Recognition	Ar, Hr, Ie, Sd, Ss	Natural
Scherer (2000a)	German	4 Actors	-	Ecological	Ar, Dt, Fr, Jy, Sd	Simulated
Scherer et al. (2002)	English, German	100 Native	-	Recognition	2 Tl, 2 Ss	Natural
Schiel et al. (2002), SmartKom	German	45 Native	V	Recognition	Ar, Dfn, Nl	Natural
Schröder and Grice (2003)	German	1 Male	-	Synthesis	Soft, modal, loud	Simulated
Schröder (2000)	German	6 Native	-	Recognition	Ar, Bm, Dt, Wy, ...	Simulated
Slaney and McRoberts (2003), Babyears	English	12 Native	-	Recognition	Al, An, Pn	Natural
Stibbard (2000), Leeds	English	Unknown	-	Recognition	Wide range	Natural, elicited
Tato (2002), AIBO	German	14 Native	-	Synthesis	Ar, Bm, Hs, Nl, Sd	Elicited
Tolkmitt and Scherer (1986)	German	60 Native	-	Recognition	Cognitive Ss	Elicited
Wendt and Scheich (2002), Magdeburger	German	2 Actors	-	Recognition	Ar, Dt, Fr, Hs, Sd	Simulated
Yildirim et al. (2004)	English	1 Actress	-	Recognition	Ar, Hs, Nl, Sd	Simulated
Yu et al. (2001)	Chinese	Native from TV	-	Recognition	Ar, Hs, Nl, Sd	Simulated
Yuan (2002)	Chinese	9 Native	-	Recognition	Ar, Fr, Jy, Nl, Sd	Elicited

Abbreviations for emotions: The emotion categories are abbreviated by a combination of the first and last letters of their name. At: Amusement, Ay: Antipathy, Ar: Anger, Ae: Annoyance, Al: Approval, An: Attention, Anxty: Anxiety, Bm: Boredom, Dfn: Dissatisfaction, Dom: Dominance, Dn: Depression, Dt: Disgust, Fd: Happiness, Ie: Indifference, Iy: Irony, Jy: Joy, Nl: Neutral, Pc: Panic, Pn: Prohibition, Se: Surprise, Sd: Sadness, Ss: Stress, Sy: Shyness, Sk: Shock, Td: Tiredness, Tl: Task load stress, Wy: Worry. Ellipses denote that additional emotions were recorded.

Abbreviations for other signals: BP: Blood pressure, BL: Blood examination, EEG: Electroencephalogram, G: Galvanic skin response, H: Heart beat rate, IR: Infrared Camera, LG: Laryngograph, M: Myogram of the face, R: Respiration, V: Video.

Other abbreviations: H/C: Hot/cold, Ld eff.: Lombard effect, A-stress, P-stress, C-stress: Actual, Physical, and Cognitive stress, respectively, Sim.: Simulated, Elic.:Elicited, N/A: Not available.

B ADDITIONAL RESULTS

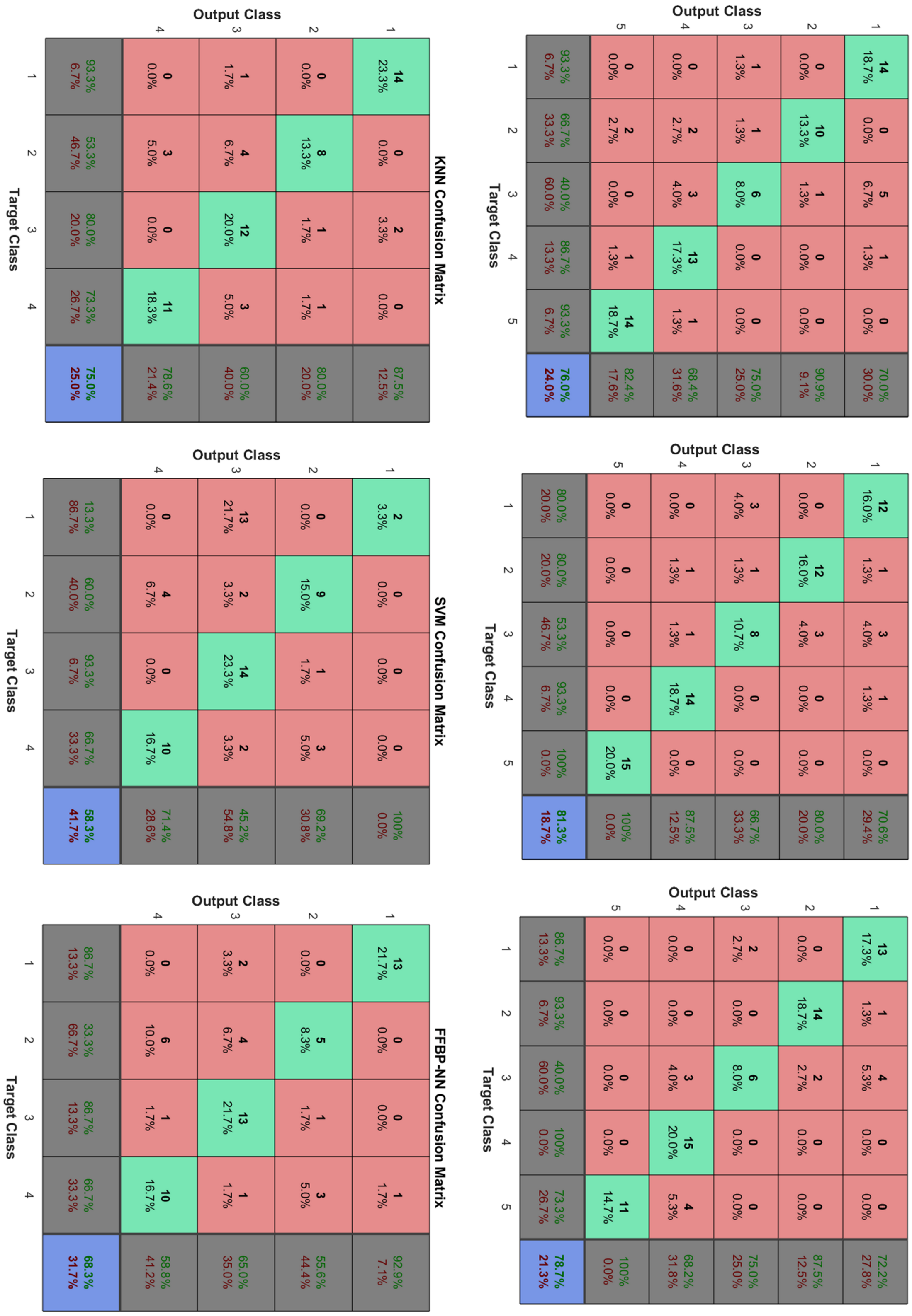


Fig. B.1: Confusion matrix of classifiers on BerlinDB (5 classes) and emoDBova (4 classes).

CONTENT OF FLASH DRIVE

At the end of the thesis is attached flash drive that contains:

- database - all used databases
- source code - *m*-files and other supported files
- manual
- other files - presented Ph.D. thesis in TEX version