BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF ELECTRICAL ENGINEERING
AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

Ing. Hicham Atassi

EMOTION RECOGNITION FROM ACTED AND SPONTANEOUS SPEECH

ROZPOZNÁNÍ EMOČNÍHO STAVU Z HRANÉ A SPONTANNÍ ŘEČI

SHORTENED VERSION OF PH.D. THESIS

Specialization: Teleinformatics
Supervisor: Prof. Ing. Zdeněk Smékal, Csc.
Opponents:

Date of Defense:

KEYWORDS

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

KLÍČOVÁ SLOVA

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

Contents

1. Introduction

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

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

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

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

The thesis aims can be summarized as follows

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

2. Emotion recognition from acted speech

2.1 Emotion recognition using BDES

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

2.1.1 Database description

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

2.1.2 Feature extraction and selection

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

2.1.3 Classifier selection

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

By using segmental features reported in the previous section, the GMM classifier outperformed the other classifiers and hence it was selected for further experiments. The results achieved for different classifiers and feature extraction schemes are illustrated in Figure 2.1.

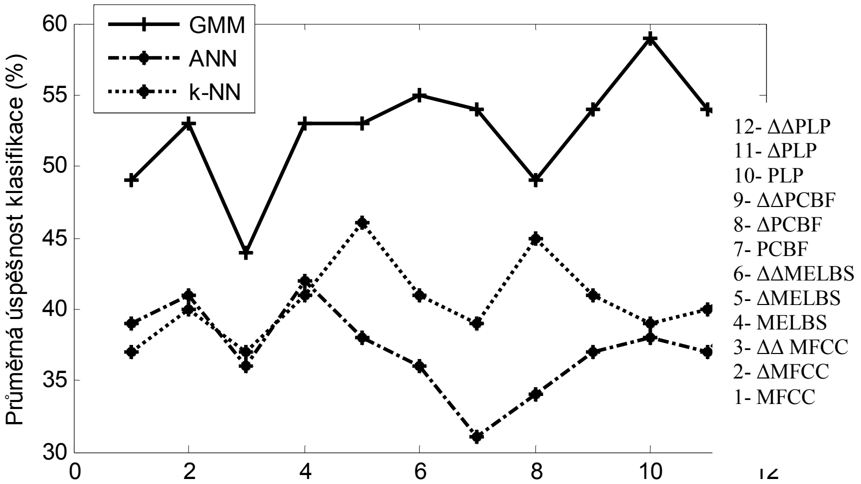

Figure 2.1: Classification accuracy for different classifiers and feature extraction schemes for BDES database.

Figure 2.2 displays the mean classification rates (averaged over the six emotions under examination) obtained using the encoding procedures and the GMM classifier discussed above (bars 1 up to 13). The best result (63%) was obtained through a combination of PLP, ΔPLP, PCBF and ΔΔMELB coefficients. The features included in this combination were identified using the Sequential Floating Forward Selection (SFFS) algorithm.

The details of the obtained results are displayed by the confusion matrix in Table 2.1. It can be noticed that satisfying classification rates are obtained only for anger (92%) and sadness (76%) emotional states. It can also be noticed that happiness emotional state is highly confused with an anger one, as well as a bored emotional state is confused with a neutral one. The high confusion between specific couple of emotional states suggests the need for a further processing and classifying step which should provide a way to overcome such problems. Moreover, since the GMM output is a likelihood valued, it should make sense to take into account couples of emotional states that have obtained from the first classifying step the highest likelihoods.

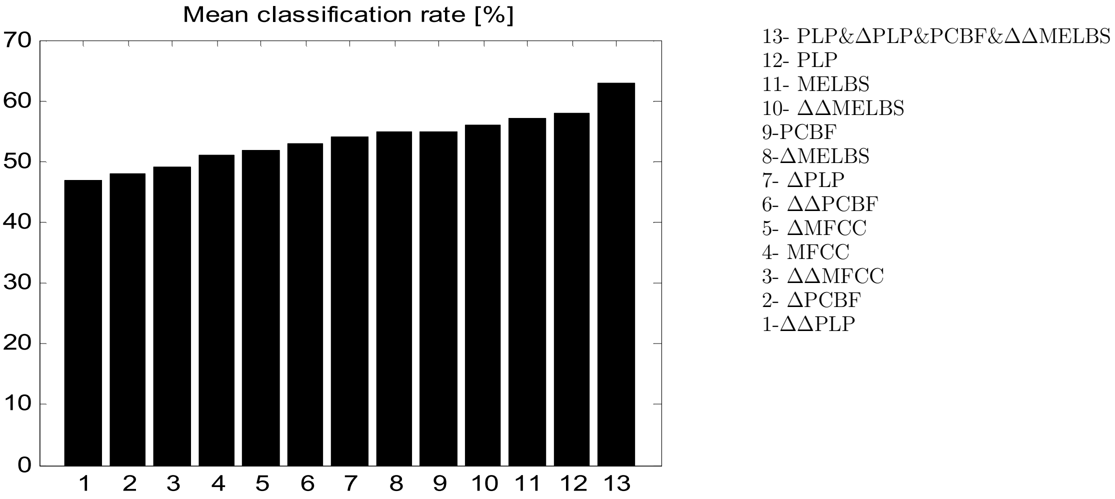

Figure 2.2: Mean classification rates for different perceptual features.

Table 2.1: Confusion matrix obtained within the first step (average classification rate: 63%)

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

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

2.1.4 Feature extraction within the second step

In the second step new feature extraction techniques were applied to the emotional speech utterances in order to identify for each couple of emotions a unique set of acoustic emotional features capable of discriminating between them. To this aim, prosodic and voice quality features including pitch, energy, zero crossing rate and harmonicity. These features were extracted from speech frames 125 ms long with a 50% of overlap. In this further processing step, the frame length was reduced in order to obtain acoustic vectors described by more prosodic and voice quality details. A total of 72 high-level features were obtained, those that revealed to be relevant in the discrimination of the 15 couples of emotional states under examination. The cross-emotion classification rates between each couple are reported in Table 2.2.

Table 2.2: Cross-emotion recognition within second classification step (average classification rate: 95.7 %).

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

Combining the two processing steps by considering the couple of emotions with the highest likelihoods in the first step as an input of the second step, the final confusion matrix obtained is shown in Table 2.3. Figure 2.3 illustrates the mechanism of the introduced approach on an example. It is worth to note that, even though at the output of the first step, the neutral emotional state obtained a highest likelihood score, boredom (which was the emotional state in input), was correctly classified as output of the second step.

Table 2.3: The final confusion matrix for BDES (average classification rate: 80.7%).

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

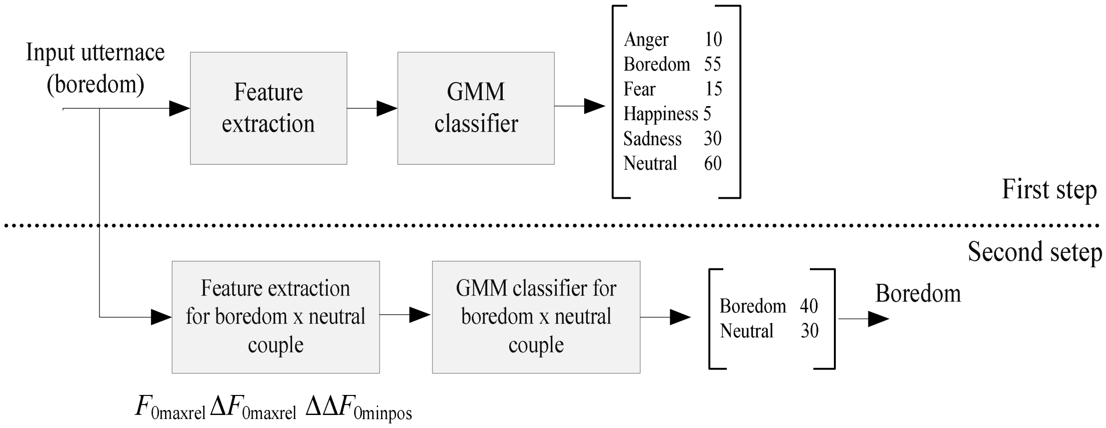
Input utternace (boredom)
Feature extraction
GMM classifier
Anger 10
Boredom 55
Fear 15
Happiness 5
Sadness 30
Neutral 60
First step
Second setep
Feature extraction for boredom x neutral couple
GMM classifier for boredom x neutral couple
Boredom 40
Neutral 30
Boredom
$F_{0\mathrm{maxrel}}$ $\Delta F_{0\mathrm{maxrel}}$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

Figure 2.3: An example of automatic vocal emotion recognition performed by "six by two approach".

2.1.5 Summary

This section proposed a new approach for automatic speaker-independent vocal emotion recognition. The main idea is in the splitting of the recognition task into two steps. The first step implements a coarse encoding and classification of the six emotional states under examination in order to identify among them the couple with the highest likelihoods. For this step, segmental-based representation of features were used and the features extracted were combined in order to obtain acoustic vectors best descriptive of emotional states. The second steps re-encode, through a set of prosodic and voice quality suprasegmental features, the emotional states that obtained the highest likelihood scores in the first step. The average classification rate obtained (80.7 %) by the proposed approach was better than the result (74.5%) reported by the most recent work of Lugger &Yang on speaker-independent recognition of emotional vocal expressions.

2.2 Emotion recognition using CDES

2.2.1 Database description

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

2.2.2 Feature extraction

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

2.2.3 Feature reduction

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

If each utterance is considered as a single input pattern in the classification process, it should be taken into account the different number of segments obtained from each utterance, that yields feature vectors with various lengths after the concatenation of the spectral characteristics (such as MFCC) extracted from the frames of each utterance under examination. A possible solution to this inconvenient is to zero padding in order to have vectors of the same length. However, this solution didn't give satisfying results in terms of classification accuracy. A better way to handle the problem would be to use feature space reduction techniques such as Vector Quantization algorithms like *k*-means or Principal Component Analysis (PCA). In the present approach a relatively simple space reduction method is proposed based on spectral vector averaging that had provided better results than PCA or *k*-means on the data under examination. The principle of the proposed space reduction approach called Temporal Mean Vector (TMV) is shown in Figure 2.4 and is applied according to the following steps:

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

$$\mathbf{x}_m^j = \frac{\sum_{n=\alpha_1}^{\alpha_2} \mathbf{x}_n^j}{\left\lfloor \frac{N}{3} \right\rfloor}, \qquad j = 1, 2, 3. \tag{2.1}$$

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

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

$$\mathbf{x}_{\mathrm{sp}} = [\mathbf{x}_m^1, \mathbf{x}_m^2, \mathbf{x}_m^3]^{\mathbf{T}} \tag{2.2}$$

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

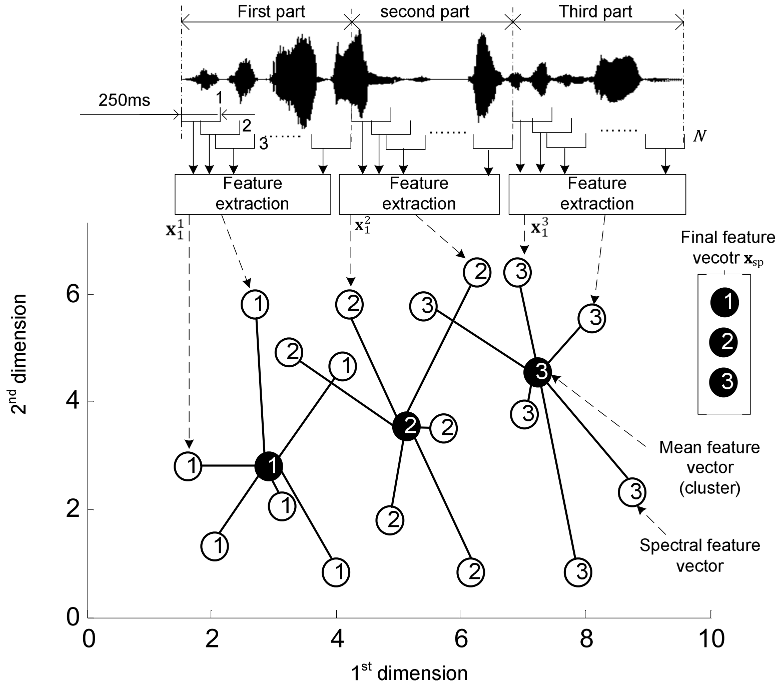

Figure 2.4: The principle of Temporal Mean Vector Method for feature reduction.

The advantage of the SFFS algorithm is that it identifies the best features according to their classification accuracy by using an arbitrary classifier. In own experiments, a GMM classifier (with one Gaussian per class) was used both for feature selection via the SFFS algorithm and for the overall validation of the proposed vocal emotion recognition algorithm. Thus, it could be stated that the features were chosen with a high level of reliability since each GMM classifier used in the final proposed system has already found its optimal features through the SFFS algorithm. The features that showed the best performance in distinguishing within couples of emotions are reported in Table 2.4.

2.2.4 Classification

The proposed classifier uses two classification techniques fused together in two classification steps as illustrated in Figure 2.6.

The first step (Figure 2.5) is used to train each sub-classifier $D^{(i)}$ to distinguish within couples of emotions. The likelihoods output by each classifier are then multiplied by each other. The final decision about the classes scores $\aleph_\omega$, $\aleph_{\omega'}$ is made according to the following formula

$$\aleph_\omega = \begin{cases} 1 & \text{for} \quad P(\omega|\mathbf{x}_{\mathrm{sp}})P(\omega|\mathbf{x}_{\mathrm{pv}}) > P(\omega'|\mathbf{x}_{\mathrm{sp}})P(\omega'|\mathbf{x}_{\mathrm{pv}}) \\ 0 & \text{for} \quad P(\omega|\mathbf{x}_{\mathrm{sp}})P(\omega|\mathbf{x}_{\mathrm{pv}}) < P(\omega'|\mathbf{x}_{\mathrm{sp}})P(\omega'|\mathbf{x}_{\mathrm{pv}}) \end{cases}, \tag{2.3}$$

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

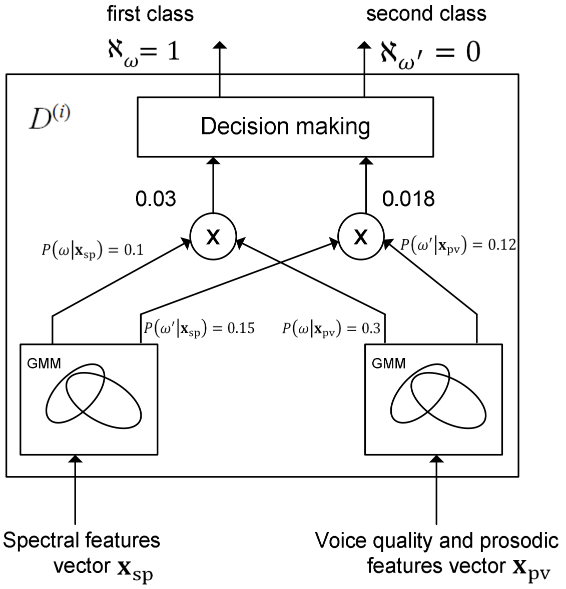
first class
second class
$\aleph_{\omega} = 1$
$\aleph_{\omega'} = 0$
$D^{(i)}$
Decision making
0.03
0.018
X
X
$P(\omega|\mathbf{x}_{\mathrm{sp}}) = 0.1$
$P(\omega'|\mathbf{x}_{\mathrm{pv}}) = 0.12$
$P(\omega'|\mathbf{x}_{\mathrm{sp}}) = 0.15$
$P(\omega|\mathbf{x}_{\mathrm{pv}}) = 0.3$
GMM
GMM
Spectral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

Figure 2.5*:* The fusion of prosodic-voice quality and spectral feature vectors.

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

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

The neurons have a linear transfer function described by

$$y_{\omega} = \sum_{i=1}^{N} \aleph_{i}^{\omega}, N = 6. \qquad (2.4)$$

2.2.5 Results

The proposed algorithm was validated using the leave-one-speaker-out validation technique. The classification rates within couples of emotions are reported in Table 2.6. The final confusion matrix is shown in Table 2.7 and the average classification rate was 60.7%. Figure 2.7 illustrates the differences in accuracy between the automatic and human subjective classification. The correlation between them is high (the normalized correlation here is R=0.79) even though the proposed system performs better for all the emotions except irony.

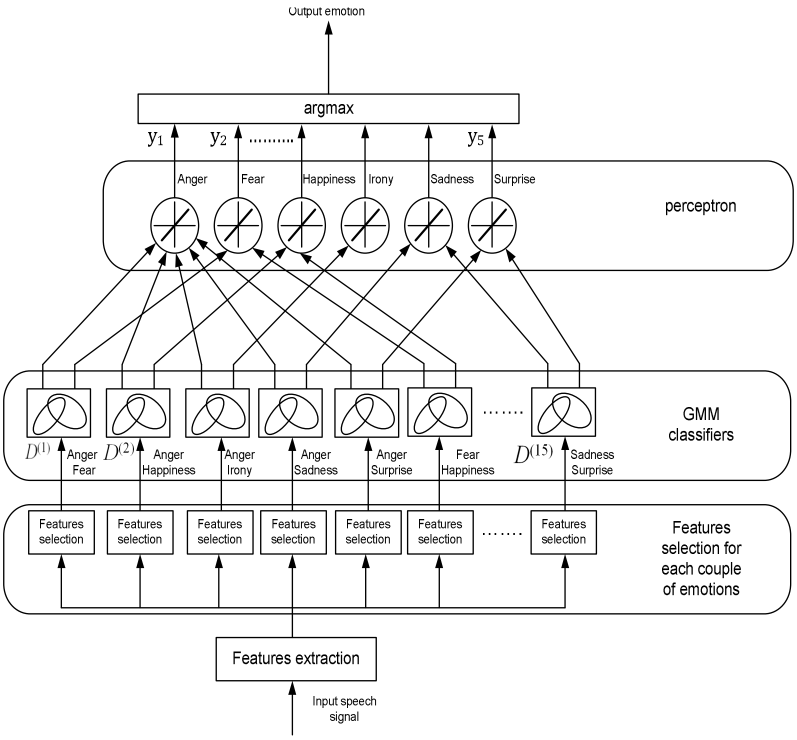

Figure 2.6: Proposed classifier based on emotion coupling.

Table 2.6: Cross-emotion recognition within couples of emotions (average classification rate: 76.4 %).

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

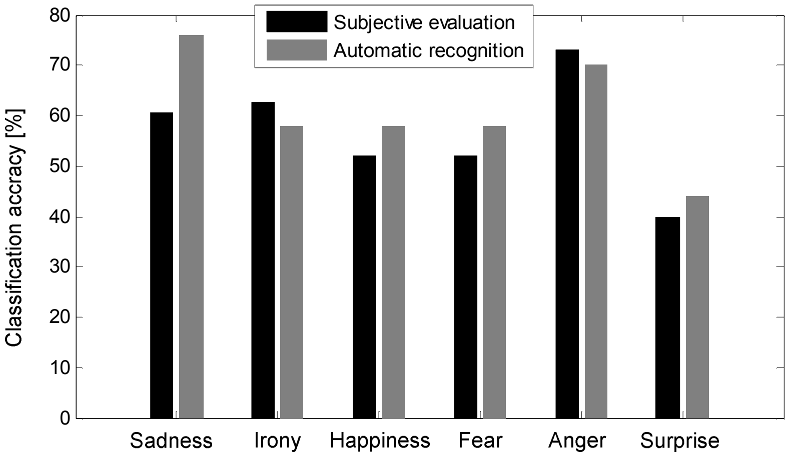

Figure 2.7: Classification performance achieved by the proposed system and by human subjects.

Table 2.7: The final confusion matrix for the six emotions under examination (average classification rate: 60.7 %).

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

2.2.6 Summary

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

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

3. Emotion recognition from spontaneous speech

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

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

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

3.1 Regression based emotion recognition from Slavic speech

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

3.1.1 Feature extraction

Feature extraction process is done as follows:

1- Speech signal is segmented into frames of 32ms with 50% overlap.
2- Features in Table 3.1 are extracted from each frame.
3- High-level characteristics are computed from segmental features obtained from the previous step.
4- High-level characteristics are concatenated into the final feature vector used for training.

Beside the features listed in Table 3.1, the first and second differences (Δ, $\Delta\Delta$) of these features were also considered.

Table 3.1: List of features for emotion recognition from Slavic speech.

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

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

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

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

Table 3.2: List of suprasegmental (high-level) features extracted from segmental features for emotion recognition from Slavic speech.

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

3.1.2 Feature selection

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

3.1.3 Regression

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

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

The speech corpus was split into three parts, where 80% of this corpus was used for training, 10% for validation and 10% for testing. The mean absolute errors for all regression techniques are reported in Table 3.3, where the maximum possible error for each emotion is 10, which corresponds to the number of listeners who were involved in the subjective evaluation process for Czech, Polish and Russian. The subjective evaluation results for Slovakian language were doubled in order to have all emotions in the same scale.

Table 3.3: Mean absolute errors for regression algorithms under examination.

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

The results suggests that the support vector regression with radial basis kernel combined with feature forward selection (SVR-RBF-FS) gives the best result among the other algorithms. Hence, the next is devoted to give more details about SVR-RBF-FS method. The MAEs for all emotions using this method are shown in Figure 3.1.

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

$$E_{\mathrm{mae}} = \frac{1}{MN} \sum_{c=1}^{M} \sum_{i=0}^{N-1} |(s_c^i - y_c)| \quad , \tag{3.2}$$

where:

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

The forward feature selection is described in the following pseudocode

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

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

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

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

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

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

        ENDFOR  // end of validation cycle

    ENDFOR // end of feature adding cycle

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

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

3.1.4 Mapping emotions onto two-dimensional space

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

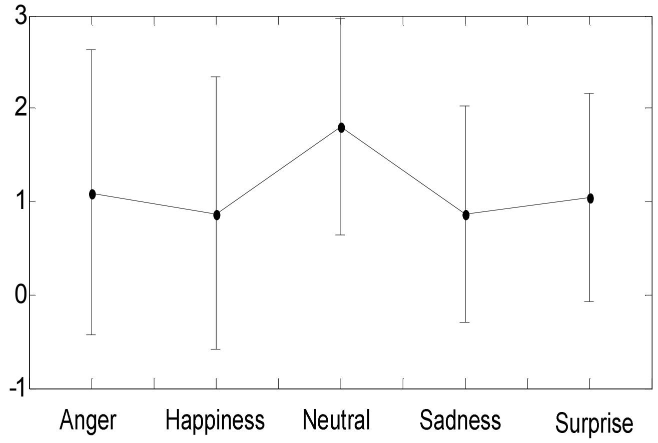

Figure 3.1: Mean Absolute errors for SVR-RBF-FS algorithm.

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

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

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

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

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

$$\mathrm{y} = C\left(\mathbf{F}\left(Z_{\mathrm{opt}}\right)\right) \tag{3.1}$$

Ratios d_{ij} defined as (3.2) are computed for all possible pairs of emotional states, these ratios determine the positions of points b_{ij} in the two dimensional emotional space. These points are always located on the connecting lines between what is called “fixed points”. For example, b_{13} is located between anger and neutral and b_{25} is located between happiness and sadness.

$$d_{ij} = \frac{y_i}{y_i + y_j}, \tag{3.2}$$

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

The fixed points represent the positions of emotions considered in the experiment; these positions were set according to the activation-valence theory proposed in [MR74]. The fixed point coordinates are shown in Table 3.4.

Table 3.4: The coordinators of fixed points.

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

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

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

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

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

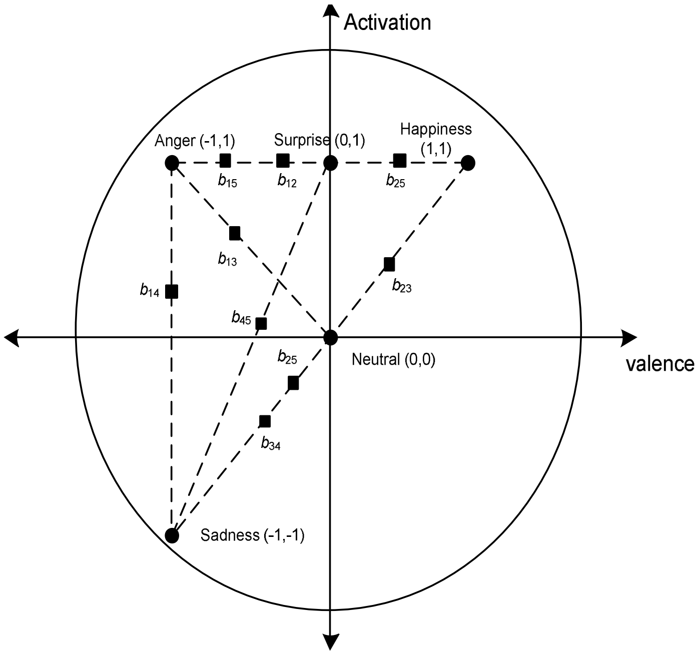

Figure 3.2: Two-dimensional emotional space with fixed and *b* points.

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

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

The remaining ratios are computed analogously:

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

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

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

Again, the remaining points are taken analogously:

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

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

The secondary evaluation was carried out by 5 listeners, who labeled a small corpus of 50 emotional speech utterances according to activation-valence protocol. These listeners were asked to guess the position of speakers' emotional states from 50 utterances. The same utterances were analyzed by the proposed algorithm and the results of subjective evaluation and automatic recognition were compared in terms of MAE. The mean absolute errors for valence and activation obtained by the two-dimensional approach are reported in Table 3.5. The maximum possible error here is 2, which the difference between the minimum (-1) and maximum (+1) values of valence and activation.

Table 3.5: Mean absolute errors for both valence and activation using skeleton approach.

Valence	0.37
Activation	0.39

As a practical example, Figure 3.3 illustrates the results of emotional analysis over client-operator phone call; the first column illustrates the two-dimensional emotional space of each channel whereas the second column represents the activation and valence values in time domain. The big advantage of the two-dimensional representation is that it gives a very good overview of the speaker's emotions distribution within the utterance as well as the intensities of these emotions. However, the basic form of such interpretation doesn't contain any temporal information, that is, it is not possible to define when a concrete emotion occurs. This limitation can be avoided by displaying the activation and valence in time (like the right column of Figure 3.3) or by using a three-dimensional interpretation, where the third dimension is time.

3.1.5 Summary

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

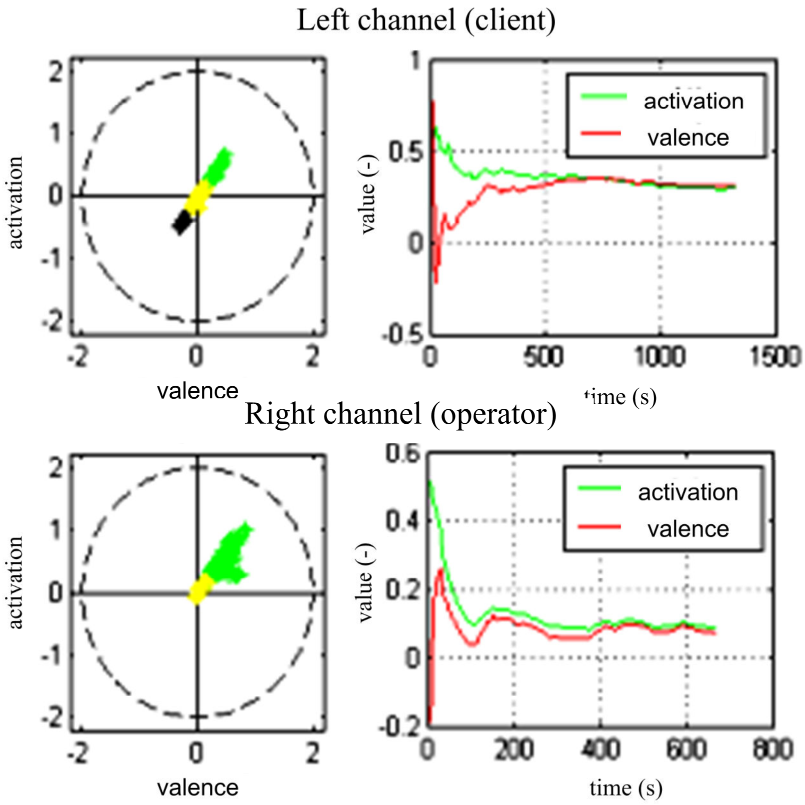

Figure 3.3: Results of emotional analysis over client-operator phone call.

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

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

3.2.1 Adaptation of Speech database

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

- The fear state was discarded due to the lack of sufficient speech with this emotion and moreover, since the emotion recognition system will be used in call centers which focus on telemarketing and customer care services, there was no interest in this emotion.
- It was found that the surprise state can't be considered as an independent state, because speakers can express surprise in both positive and negative way. Thus it will be processed independently.
- The anger and happiness classes were split into two classes; one with low activation level and one with high activation level. The anger with low activation level can approximately represent annoy emotional state where as the low level happiness represents the please emotion.

The numbers of utterances used in further experiments for each state are reported in Table 3.6.

Table 3.6: Number of utterances used for emotion recognition using MSDES.

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

3.2.2 Preprocessing and feature extraction

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

ACW and LPCC features didn't show any significant discrimination power within emotions for both acted and spontaneous databases and thus they will not be considered in the following experiments. Moreover, despite that Wavelet Decomposition WADE showed a good performance in terms of distinguishing between emotional states. It was decided to discard them as well due to their computational complexity. The features considered for emotion recognition using MSDES are reported in Table 3.7.

Table 3.7: List of features used for emotion recognition using MSDES.

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

3.2.3 General classifier

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

Two methods for fusing segmental and suprasegmental features are proposed (Figure 3.4)

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

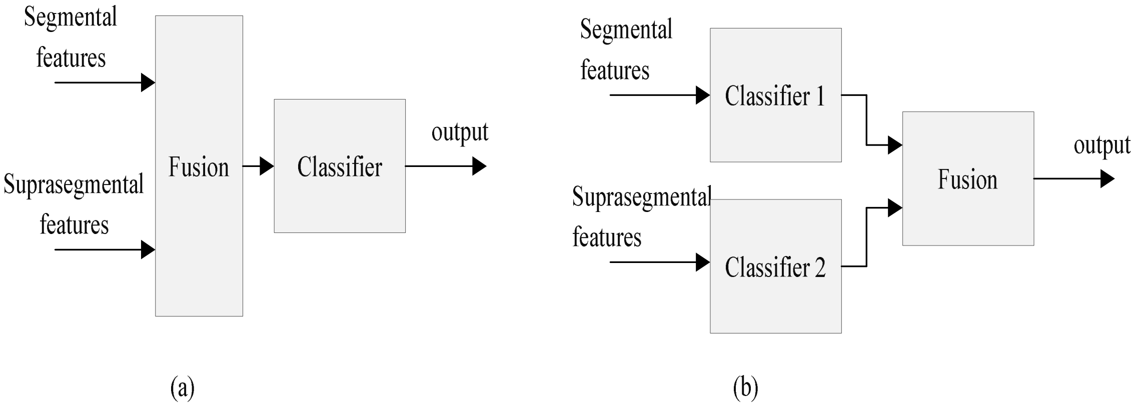

Figure 3.4: Two possible ways of fusing segmental and suprasegmental features. Fusion on feature level (a) and fusion on classifier level (b).

Table 3.8 contains the results of different classifiers and feature selection methods for suprasegmental features whereas Table 3.9 reports the results for segmental features. The classification accuracies obtained by fusing both types of features are shown in Table 3.10.

Table 3.8: Weighted Classification accuracies using suprasegmental features.

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

Table 3.9: Weighted Classification accuracies using segmental features.

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

Table 3.10: Classification results for different types of fusion of segmental and suprasegmental features.

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

Results in tables 3.8 and 3.9 suggests that support vector machines classifier with features selected by combining mRMR and SFFS algorithm gives best classification accuracy (62.45%) for suprasegmental features. Regarding segmental features, the best result is achieved by applying Temporal Mean Vector (TMV) for feature reduction and again, SVM-RBF classifier shows the best performance.

The fusion results reported in Table 3.10 indicate that there is a slight improvement when the feature vectors are fused on feature level. On the other hand, the fusion on classifier level yields decreasing the classification accuracy comparing to the case when only suprasegmental features are used.

3.2.4 Gender-dependent system

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

Table 3.11: Weighted Classification accuracies for male speakers using suprasegmental features.

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

Table 3.12: Weighted Classification accuracies for male speakers using segmental features.

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

Table 3.13: Classification results for different types of fusion of segmental and suprasegmental features for male speakers.

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

Table 3.14: Classification accuracies for female speakers using suprasegmental features.

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

Table 3.15: Classification accuracies for female speakers using segmental features.

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

Table 3.16: Classification results for different types of fusion of segmental and suprasegmental features for female speakers.

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

For gender-dependent approach, the results reported in Tables 3.11 to 3.16 proves that the combination of mRMR and SFFS algorithms for feature selection and SVM-RBF classifier works best for both male and female speakers. The fusion of segmental and suprasegmental features didn't reveal any significant improvement comparing to the case when only suprasegmental features are used. Thus it was decided to discard segmental features in further experiments.

3.2.5 Emotion coupling

Despite that this approach was validated on acted speech only with GMM classifiers, an arbitrary classifier can be used as a classification element in this approach. The first task then is to find the best classification element for emotion coupling that best fits MSDES. Table 3.17 reports results of four different classifiers using suprasegmental features and mRMR for feature selection. The results suggest that SVM-RBF classifier works best for emotion coupling. The confusion matrices for SVM-RBF with mRMR and SVM-RBF with mRMR+SFFS are reported in Tables 3.18 and 3.19 respectively.

Table 3.17: Classification accuracies for different classification elements of emotion coupling approach.

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

Table 3.18: Confusion matrix for emotion coupling using SVM-RBF with mRMR (average classification accuracy: 61.8096%)

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

Table 3.19: Confusion matrix for emotion coupling using SVM-RBF with mRMR (average classification accuracy: 67.4135%).

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

3.2.6 Fusion of all systems

Outputs from all classifiers namely general, emotion coupling and gender-dependent are fused using one layer perceptron with 50 neurons. The outputs of the fusion layer are connected to the two-dimensional mapping layer except the surprise which is, as it was stated before, processed independently. The final confusion matrix obtained by fusing all systems is reported in Table 3.20 and the comparison of results for several feature extraction, reduction and classification techniques for MSDES are summarized in Figure 3.5. The block scheme of the complex classification system is illustrated in figure 3.6.

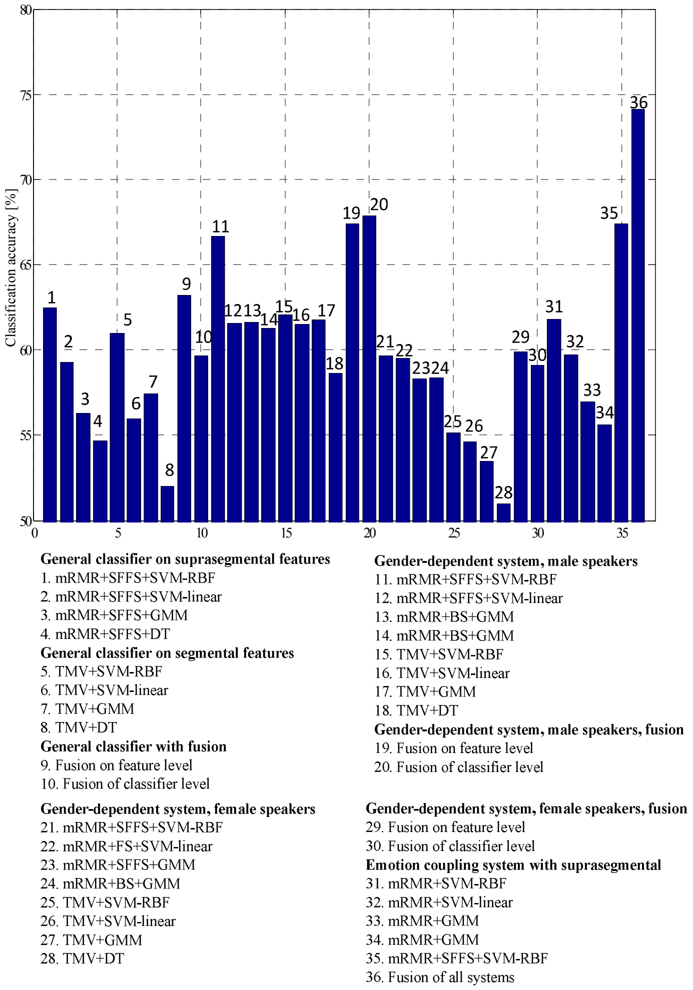

Figure 3.5: Comparison of results for several feature extraction, reduction and classification techniques for MSDES.

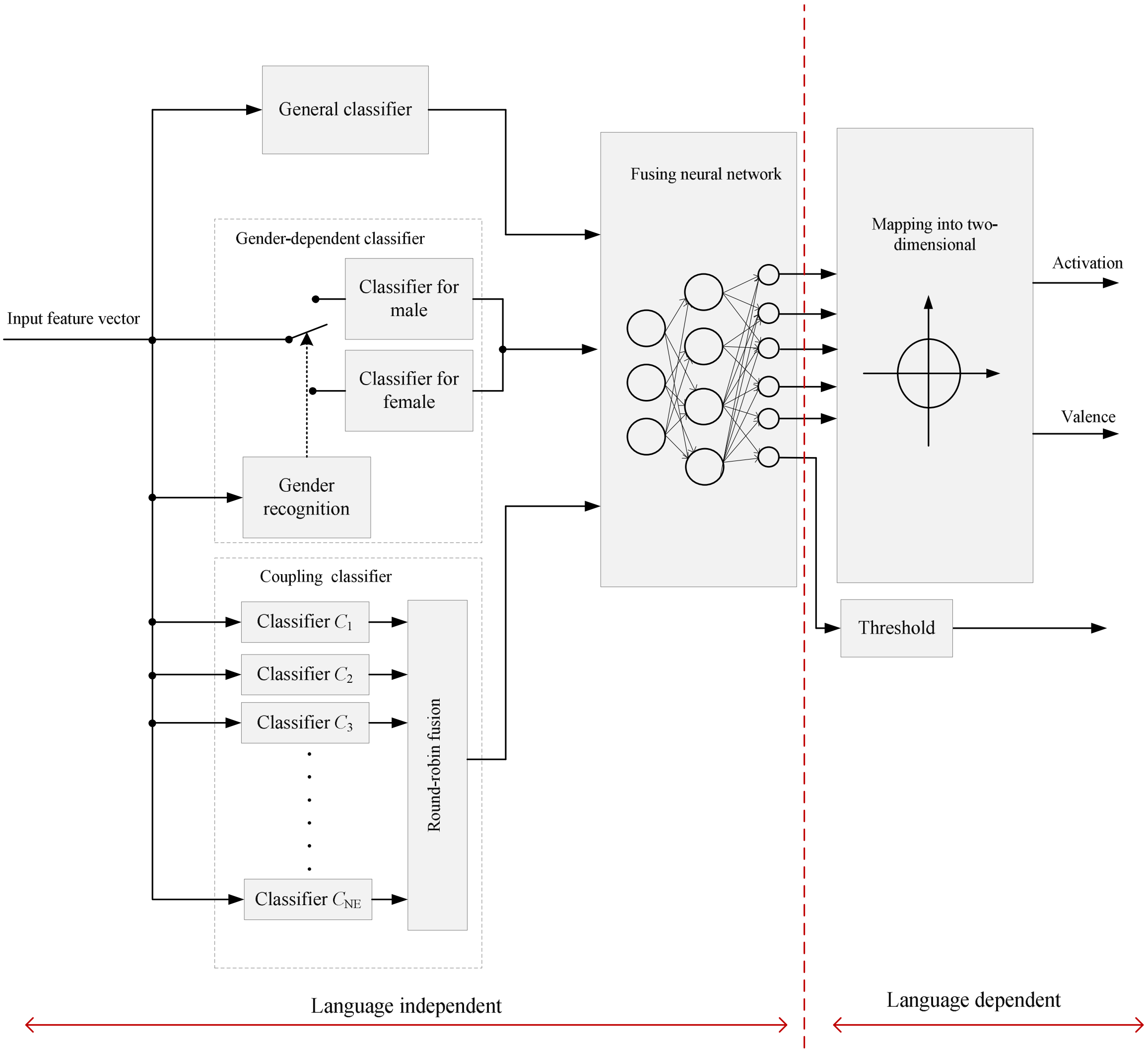

Figure 3.6: Block scheme of proposed system for emotion recognition form MSDES.

Table 3.20: Final confusion matrix obtained after fusing all classifiers (average classification accuracy 74.16%).

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

3.2.7 Mapping emotions into two-dimensional space

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

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

Table 3.21: Proposed fixed points of skeleton approach for MSDES.

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

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

The results of two-dimensional mapping in terms of mean absolute errors for both skeleton and neural network approaches are reported in Table 3.22.

Table 3.22: Evaluation of different approach for two-dimensional mapping in terms of MAE.

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

3.2.8 Detection of surprise state

The detection of surprise state is carried out by using SVM classifier with linear function, as it showed the best classification performance among other classifiers. For each supersegment, this classifier is applied, if the classifier output exceeds a certain threshold, then surprise is registered beside the original emotion. Figure 4.1 illustrates ROC curves of several classifiers for the detection of surprise state whereas the detailed results of these classifiers are reported in Table 3.23.

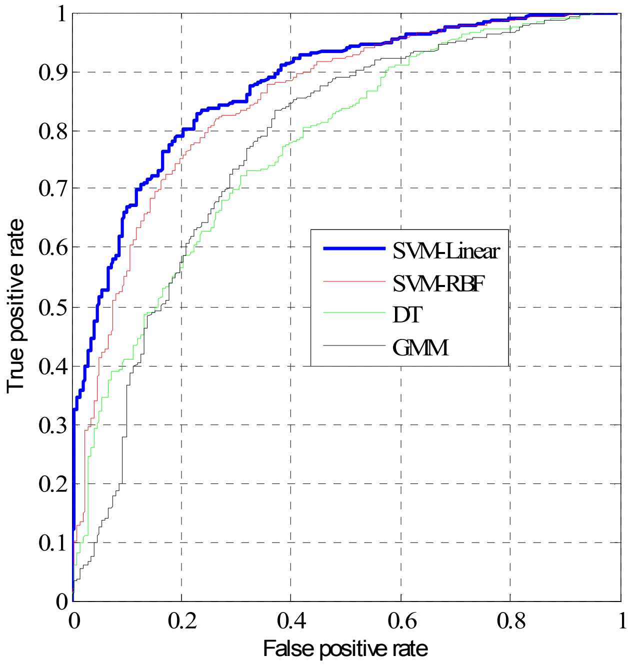

Figure 4.1: ROC curves of different classifiers for the detection of surprise state.

Table 3.23: Results of surprise detection for several classifiers.

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

3.2.9 Comparison of own system with relevant systems in literature

I believe that it is not possible to make a fair comparison of systems for emotion recognition when different speech databases are used. This is due to the variability in the number of emotions and their type, number of speakers, the matter of speech (clean/noisy) and the authenticity of emotional content. Despite of this, a brief comparison of own system with systems proposed in literature for emotion recognition in the domain of call centers are reported in Table 3.24.

Table 3.24: Comparison of own system with systems reported in literature.

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

The comparison reported in previous Table reveals that high classification accuracy of spontaneous emotions in the domain of call centers are performed with high accuracy when only two emotional states (anger and neutral) are considered. However, own approach works with 6 different emotional states. In case that only two states namely high anger and neutral are taken into account from the six reported in Table 3.22, then the proposed system outperforms approaches presented in [Yac+03, CD11, YP07], giving classification accuracy of 97.5%. The systems presented in [Lau+11, VS11] works with 3 and 4 emotional states respectively. Again, the proposed system significantly outperforms these mentioned systems.

3.2.10 Discussion

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

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

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

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

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

The aim of the research proposed in this chapter [AS14] is to investigate the possibility of using dialogue features obtained from agent-client conversations to automatically identify successful phone calls in call centers. This can be very handy to spot the unsuccessful sessions within the large database of recorded telephone calls and can help the supervisors of call centers to figure out mistakes made by their agents. The basic block scheme of the proposed approach is illustrated in Figure 4.1and the next sections are devoted to presenting each block in details.

4. Analysis of spoken dialogue

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

approach is illustrated in Figure 4.1 and the next sections are devoted to presenting each block in details.

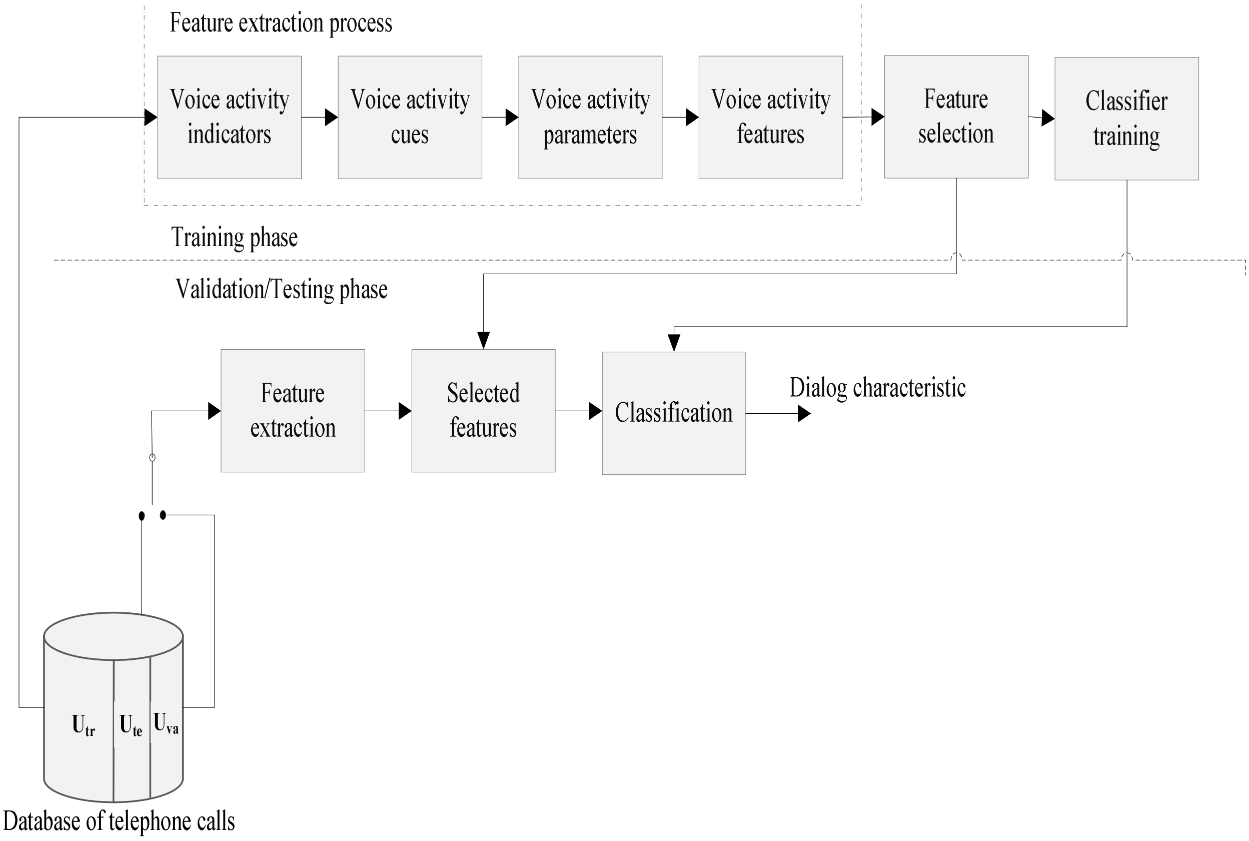

Figure 4.1: Basic block scheme of presented approach for successful call identification.

4.1 Description of speech corpus

The speech corpus used in experiments contains 48 dual-channel agent-client call records obtained from call centers in the Czech Republic. The basic characteristics of this corpus are reported in Table 4.1. It is worth mentioning that there is no statistical difference between the duration average values of successful and unsuccessful calls (p=0.29).

Table 4.1: Basic characteristics of speech corpus, the average duration of records and their standard deviation are in seconds.

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

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

4.2 Feature extraction

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

4.2.1 Extraction of voice activity indicators

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

4.2.2 Extraction of voice activity cues

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

- **Hesitations**: silence periods or filled pauses that occur within the voice activity waveform in one direction that are not followed by voice activity in the other direction.
- **Reactions** (or turn takings): A reaction is registered when uninterrupted voice activity in one direction is followed by voice activity in the other direction.
- **Interruptions** (or overlaps): An interruption is registered when voice activity occurs simultaneously in both channels. The cues mentioned above are graphically illustrated on Figures 4.2.
- **Cumulative voice activity (CVA)**: this cue is proposed as a simple indicator of voice activity distribution over time. The formula that describes CVA is as follows

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

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

4.2.3 Extraction of voice activity parameters

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

$H_p = (64000, 328000, 562000)$,

$H_L = (4000, 8000, 2500)$.

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

4.2.4 Extraction of voice activity features

Statistical features are computed from vectors extracted in previous step and from CVA as well, these features are subsequently concatenated into the final feature vector used for training.

The result of feature extraction process is a vector of 490 statistical features extracted from both channels. For better understanding, the numbers of statistical features extracted from each cue/parameter are given in Table 4.2.

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

Table 4.2: The number of statistical features extracted from each cue/parameter.

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

4.3 Experiment results

After extracting features and identifying those with most discriminative power for each classifier, leave-one-out validation is performed in order to assess the performance of these classifiers. Because our two classes (successful and unsuccessful phone calls) are not equally distributed, different evaluation measurements are reported, namely weighted accuracy, unweighted accuracy, precision, recall, F-measure and Matthews correlation. Moreover, ROC curves of each classifier are shown in Figure 4.3.

The experiment results reported in Table 4.3 shows that SVM classifier with linear function performed the best in terms of precision. However, this classifier evinced low recall comparing to NBC and SVM-RBF. The last mentioned classifier had, beside NBC, perfect (100%) recall rate. Nonetheless, SVM-RBF showed the lowest precision among all classifiers. Since both recall and precision are considered as important factors, the F-measure was selected as a criterion for classifier selection as this measurement takes into accounts both recall and precision. In terms of this criterion, NBC classifier performed the best with F-measure of 96%.

Table 4.4 contains features selected using SFFS for NBC. These features are reported in descending order from the most relevant to the least relevant. It is worth mentioning here that first four features were selected within SFFS as most relevant for all classifiers under examination.

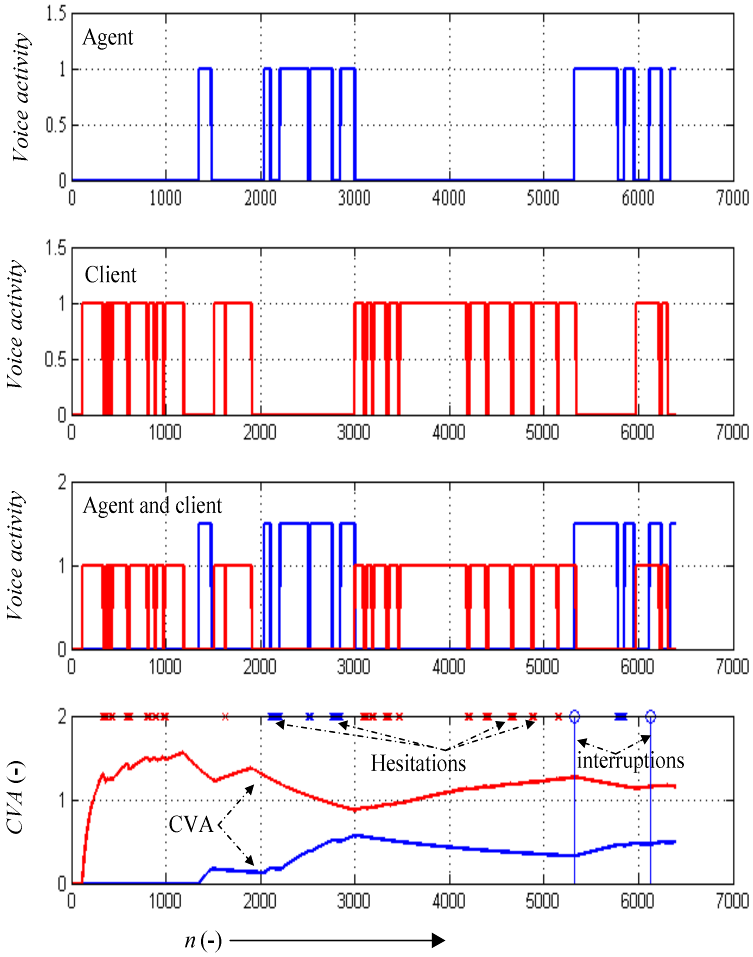

Figure 4.2: Illustration of agent's and client's voice activity indicators (up) and the corresponding cumulative voice activity cues (down).

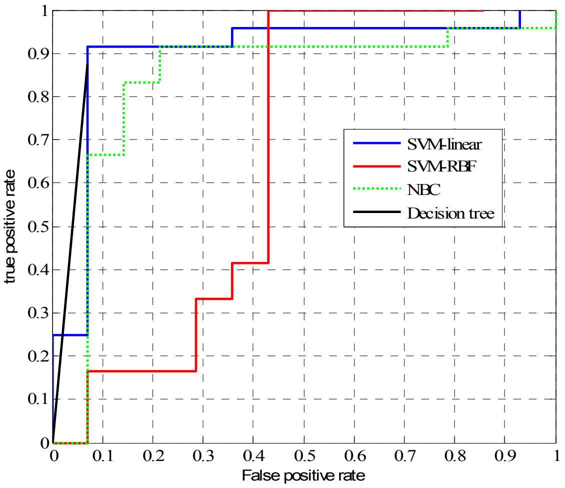

Figure 4.3: ROC curves of each classifier.

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

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

Table 4.3: Experiment results in terms of different evaluation measurements.

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

Table 4.4: Most significant dialogue features.

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

5 Conclusions

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

Regarding emotion recognition from acted speech, a new speaker independent procedure for classifying vocal expressions from Berlin Database of Emotional speech was proposed. The procedure is based on the splitting up of the emotion recognition process into two steps. In the first step, a combination of selected acoustic features is used to classify six emotions through a Bayesian Gaussian Mixture Model classifier (GMM). The two emotions that obtained the highest likelihood scores are selected for further processing in order to discriminate between them. For this purpose, a unique set of high-level acoustic features was identified using the

Sequential Floating Forward Selection (SFFS) algorithm, and a GMM was used to separate between each couple of emotion. The mean classification rate is 81% with an improvement of 6% with respect to the most recent results obtained on the same database (75%).

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Curriculum Vitae

Hicham Atassi

Personal information

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

Language Skills

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

Registered products

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

Patents

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

Publications

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

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

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

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

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

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

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

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

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

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

-KUBÁNKOVÁ, A.; ATASSI, H.; KUBÁNEK, D. Noise Robust Automatic Digital Modulation Recognition Based on Gaussian Mixture Models. In *Proceedings of the 6th International Conference on Teleinformatics - ICT 2011 (id 18951).* Brno, Czech Republic: VUT v Brne, 2011. p. 220-226. ISBN: 978-80-214-4231- 3.

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

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

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

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

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

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

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

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

Press, 2009. p. 333-336. ISBN: 978-960-474-139- 7.

-ATASSI, H.; ESPOSITO, A. A Speaker Independent Approach to the Classification of Emotional Vocal Expressions. In *Proceedings of Twentieth International Conference on Tools With Artificial Intelligence, ICTAI 2008.* Dayton, Ohio, USA: IEEE Computer Society, 2008. p. 147-152. ISBN: 978-0-7695-3440- 4.
-ATASSI, H. Fundamental frequency detection methods. *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz),* 2008, vol. 2008, no. 4, p. 1-17. ISSN: 1213- 1539.

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

Number of citations (without self-citation): 29

H-index according to Google scholar: 4

ABSTRAKT

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

KLÍČOVÁ SLOVA

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

ABSTRACT

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

KEYWORDS

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