Classification on Speech Emotion Recognition - A Comparative Study

Theodoros Iliou, Christos-Nikolaos Anagnostopoulos
Cultural Technology and Communication Department
University of the Aegean
Mytilene, Lesvos Island, GR-81100
th.iliou@ct.aegean.gr, canag@ct.aegean.gr

Abstract **- In this paper we present a comparative analysis of four classifiers for speech signal emotion recognition. Recognition was performed on emotional Berlin Database. This work focuses on speaker and utterance (phrase) dependent and independent framework. One hundred thirty three (133) sound/speech features have been extracted from Pitch, Mel Frequency Cepstral Coefficients, Energy and Formants. These features have been evaluated in order to create a set of 26 features, sufficient to discriminate between seven emotions in acted speech. Multilayer Percepton, Random Forest, Probabilistic Neural Networks and Support Vector Machine were used for the Emotion Classification at seven classes namely anger, happiness, anxiety/fear, sadness, boredom, disgust and neutral. In the speaker dependent framework, Probabilistic Neural Network reaches very high accuracy(94%), while in the speaker independent framework the classification rate of the Support Vector Machine reaches 80%. The results of numerical experiments are given and discussed in the paper.**

Keywords **-** ***Emotion Recognition , Artificial Neural Networks, Support Vector Machine, speech processing.***

I. INTRODUCTION

Recently, the information provided by cameras and microphones enable the computer to interact with the user though advanced image and sound processing techniques in systems similar to the one presented in Figure 1. Therefore, one of these skills that computer potentially can develop, is the ability to understand the emotional state of the person. In the field of human-computer interaction (HCI), emotion recognition from the computer is still a challenging issue, especially when the recognition is based solely on voice, which is the basic mean of human communication. In human-computer interaction systems, emotion recognition could provide users with improved services by being adaptive to their emotions. Therefore, emotion detection from speech could have many potential applications.

Communication is an important capability, not only based on the linguistic part but also based on the emotional part. Therefore, emotion detection from speech could have many potential applications in order to make the computer more adaptive to the user's needs [1] [2].

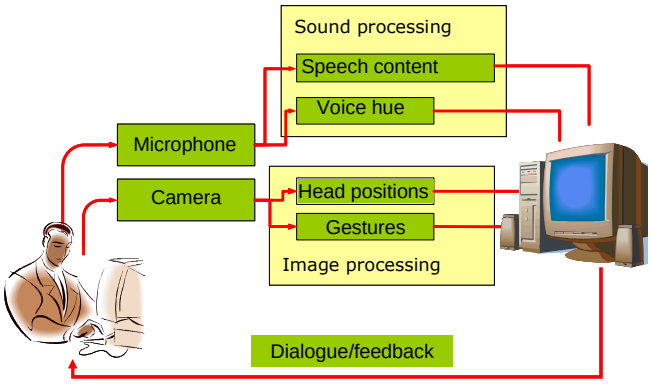

Figure 1. Human-computer interaction modules for emotion recognition.

Nowadays, with the proliferation of the Internet and multimedia, many kinds of multimedia equipment are available. Even common users can record or easily download video or audio data by himself/herself. Can we determine the contents of this multimedia data expeditiously with the computer's help? The ability to detect expressed emotions and to express facial expressions with each given utterance would help improve the naturalness of a computer-human interface. Certainly, emotion is an important factor in communication. And people express emotions not only verbally but also by non-verbal means. Non-verbal means consist of body gestures, facial expressions, modifications of prosodic parameters, and changes in the spectral energy distribution [3]. Often, people can evaluate human emotion from the speaker's voice alone since intonations of a person's speech can reveal emotions. Simultaneously, facial expressions also vary with emotions. There is a great deal of mutual information between vocal and facial expressions. Speech-driven facial animation is an effective technique for user interface and has been an active research topic over the past twenty years. Various audio-visual mapping models have been proposed for facial animation [4] [5].

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

In the computer speech community, much attention has been given to “what was said” and “who said it”, and the associated tasks of speech recognition and speaker identification, whereas “how it was said” has received relatively little. What kinds of features might carry more information about the emotional meaning of each utterance? Because of the diversity of languages and the different roles and significance of features in different languages, they cannot be treated equally [6]. It is hard to calculate, which features carry more information, and how to combine these features to get a better recognition rate.

Research in automatic detection of expressed emotion is quite limited. Recent research in this aspect mostly focuses on classification, in the other words, mostly aims at ascertaining the emotion of each utterance. This, however, is insufficient for our applications.

Various classification algorithms have been used in recent studies about emotions in speech recognition, such as k-Nearest Neighbor, NN (Neural Network), MLB (Maximum-Likelihood Bayes), KR (Kernel Regression), GMM (Gaussian Mixture Model), and HMM (Hidden Markov Model) [1] [2] [3] [7].

Previous research on emotions both in psychology and speech tell us that we can find information associated with emotions from a combination of prosodic, tonal and spectral information; speaking rate and stress distribution also provide some clues about emotions [2] [3] [8] [9]. Prosodic features are multi-functional. They not only express emotions but also serve a variety of other functions as well, such as word and sentence stress or syntactic segmentation. The role of prosodic information within the communication of emotions has been studied extensively in psychology and psycholinguistics. More importantly, fundamental frequency and intensity in particular vary considerably across speakers and have to be normalized properly [3].

Emotional inflection and modulation in synthesized speech, either through phrasing or acoustic features is useful in human-computer interaction. Such capability makes speech natural and expressive. For example a dialog system might modulate its speech to be more puerile if it deems the emotional model of its current user is that of a child. In e-learning applications, Emotion Recognition (ER) can be used to adjust the presentation style of a computerized tutor when a learner is bored, interested, frustrated, or pleased [10] [11].

Human beings are eminently emotional, as their social interaction is based on the ability to communicate their emotions and perceive the emotional states of others [4]. Designing an automatic system able to express emotions and to detect the emotional state of a user is one of the main aims of the research field defined as affective computing [4]. The acknowledgment of the user’s affective state can improve the effectiveness of a number of computer applications in a variety of fields.

Affective computing, a discipline that develops devices for detecting and responding to users’ emotions, and affective mediation, computer-based technology, which enables the communication between two or more people, displaying their emotional states [12] [5], are growing areas of research [13]. Affective mediation tries to minimize the filtering of affective information carried out by communication devices, due to the fact they are usually devoted to the transmission of verbal information and therefore, miss nonverbal information [1].

Applications of mediated communication can be textual telecommunication technologies such as affective electronic mail, affective chats, etc.

In the development of affective applications, affective resources, such as affective stimuli databases, provide a good opportunity for training such applications, either for affective synthesis or for affective recognizers based on classification via artificial neural networks (ANN), Hidden Markov Models, genetic algorithms, or similar techniques (e.g., [2] [8]). As seen in [14], there is a great amount of effort devoted to the development of affective databases. Affective databases usually record information by means of images, sounds, speech, psychophysiological values, etc. Psychological health services, i.e., counseling, benefit from affective computing applications when determining a client's emotional state. Affective computing sends a message via color or sound to express an emotional state to others. Robotic systems capable of processing affective information exhibit higher flexibility while one works in uncertain or complex environments. Companion devices, such as digital pets, use ER abilities to enhance realism and provide a higher degree of autonomy. Other potential applications are centered around social monitoring. For example, a car can monitor the emotion of all occupants and engage in additional safety measures, such as alerting other vehicles if it detects the driver to be angry. ER has potential applications in human computer interaction, such as affective mirrors allowing the user to see how he or she performs; emotion monitoring agents sending a warning before one sends an angry email; or even music players selecting tracks based on mood [10] [11]. ER is also being applied to the development of communicative technologies for use by people with autism. People with disabilities can benefit from speech emotion recognition programs. There is growing need for technologies to assist with the inhome care of the elderly and people with Alzheimer's, Parkinson's and other disabilities or traumas. Emotional speech processing recognizes the user's emotional state by analyzing speech patterns. Vocal parameters and prosody features such as fundamental frequency, intensity and speaking rate are strongly related with the emotion expressed in speech [13] [1]. However, voice quality and short-term spectral features have also to be considered when studying emotional or affective speech [2] [8] [14] and speech rate are analyzed through pattern recognition [7] [14].

In this work, we used PNN, MLP and SVM to classify seven emotions. There are a variety of temporal and spectral features that can be extracted from human speech. We used

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

statistical analysis in order to select features relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs) and Formants of speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments allow us to show, which features carry the most emotional information and which classifier reaches better accuracy.

The paper is structured as follows. The following section describes a dimensional view of emotions. Section III reports analytically the sound features that have been tested and how these features were calculated. Section IV describes the selected classifiers, as well as the two separate testing frameworks (speaker dependent/ independent recognition). The paper ends with the conclusion section, which highlights also several aspects of the emotion recognition task on the basis of sound processing.

II. BASIC EMOTIONS

According to a dimensional view of emotions, large amounts of variation in emotions can be located in a two-dimensional space, with coordinates of valence and arousal [8]. The valence dimension refers to the hedonic quality of an affective experience and ranges from unpleasant to pleasant. The arousal dimension refers to the perception of arousal associated with the experience, and ranges from very calm to very excited at the other etc. In this paper, the Berlin Emotional database (EMO-DB) [7] was used to conduct our experiments, which contains 535 utterances of 10 actors (5 male, 5 female) simulating 7 emotional states. The above set of seven emotion classes can also be well separated into two hyper classes, namely high arousal containing anger, happiness, anxiety/fear and low arousal containing neutral, boredom, disgust and sadness. The classification of disgust into low arousal can be challenged, but according to the literature disgust belongs to low arousal emotions [14]. In the experiments reported in this paper, always whole utterances have been analysed.

III. SOUND/SPEECH FEATURES

The fundamental frequency (F0), often referred to as the pitch, is one of the most important features for determining emotion in speech [3] [6] [9] [15]. Pitch represents the perceived fundamental frequency of a sound. It is one of the four major auditory attributes of sounds along with loudness, timbre and sound source location. When the actual fundamental frequency can be precisely determined through physical measurement, it may differ from the perceived pitch because of overtones, also known as upper partials, harmonic or otherwise. The human auditory perception system may also have trouble distinguishing frequency differences between notes under certain circumstances. According to ANSI acoustical terminology, it is the auditory attribute of sound according to which sounds can be ordered on a scale from low to high [16].

Bäzinger et al. argued that statistics related to pitch convey considerable information about emotional status [17]. However, pitch was also shown to be most gender-dependent feature [18]. If the recognition system ignores this issue a misclassification of utterances might be the consequence. It should be noted that most of the features that will be described below are gender-dependent to varying degrees.

Beside pitch, other commonly employed features are related to energy, speaking rate, formants, as well as spectral features, such as MFCCs. Formants are defined by Fant as the spectral peaks of the sound spectrum |P(f)|' of the voice. Formant is also used to mean an acoustic resonance and, in speech science and phonetics, a resonance of the human vocal tract. It is often measured as an amplitude peak in the frequency spectrum of the sound, using a spectrogram or a spectrum analyzer, though in vowels spoken with a high fundamental frequency, as in a female or child voice, the frequency of the resonance may lie between the widely-spread harmonics and hence no peak is visible. MFCCs are coefficients that collectively make up a mel-frequency cepstrum(MFC). In sound processing, the (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression [16].

Wang & Guan [19] and [20] used prosodic, MFCCs and formant frequency features to represent the characteristics of the emotional speech while the facial expressions were represented by Gabor wavelet features. According to Kostoulas et al. [21], an individual’s emotional state is strongly related to pitch and energy while pitch and energy of a speech signal expressing happiness or anger is, usually, higher than those associated with sadness. MFCCs have been widely used for speech spectral representation in numerous applications, including speech, speaker, gender and emotion recognition [9]. They are also increasingly finding uses in music information retrieval applications such as genre classification and audio similarity measures [22].

In this paper, pitch was extracted from the speech waveform using a modified version of the algorithm for pitch tracking proposed in [23], which is offered in the VOICEBOX toolbox [24]. Using a frame length of 100ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. If the speech is unvoiced the corresponding marker in the pitch vector was set to zero.

In addition., for each 5ms frame of speech, the first four standard MFCC parameters were calculated by taking the absolute value of the **short-time Fourier transform(**STFT), warping it to a Mel-frequency scale, taking the **discrete**

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

cosine transform (DCT) of the log-Mel spectrum and returning the first 4 components. The Matlab Code, which performs the above calculation was provided in [25]. The **STFT**, or alternatively **short-term Fourier transform**, is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. A discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. DCTs are important to numerous applications in science and engineering, from lossy compression of audio and images (where small high-frequency components can be discarded), to spectral methods for the numerical solution of partial differential equations. The use of cosine rather than sine functions is critical in these applications: for compression, it turns out that cosine functions are much more efficient (as explained below, fewer are needed to approximate a typical signal), whereas for differential equations the cosines express a particular choice of boundary conditions.

In particular, a DCT is a Fourier-related transform similar to the discrete Fourier transform (DFT), but using only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry (since the Fourier transform of a real and even function is real and even), where in some variants the input and/or output data are shifted by half a sample. There are eight standard DCT variants, of which four are common [16].

Energy, often referred to as the volume or intensity of the speech, is also known to contain valuable information. Energy provides information that can be used to differentiate sets of emotions, but this measurement alone is not sufficient to differentiate basic emotions. In [26] it is referred that fear, joy, and anger have increased energy level, whereas sadness has low energy level.

The choice of the window in short-time speech processing determines the nature of the measurement representation. The energy frame size should be long enough to smooth the contour appropriately but short enough to retain the fast energy changes, which are common in speech signals and it is suggested that a frame size of 10–20 ms would be adequate. Two representative windows are widely used, Rectangular and Hamming. The latter has almost twice the bandwidth of the former, for the same length. Furthermore, the attenuation for the Hamming window outside the pass band is much greater. Short-Time energy is a simple short-time speech measurement. It is defined by (1):

$$E_n = \sum [x(m) \cdot w(n-m)]^2 \qquad (1)$$

where m is the overlapping length of the original signal x and Hamming windowed signal w with length n. For the length of the window a practical choice is 160-320 samples (sample for each 10-20 msec) for sampling frequency 16kHz. For our experiments the Hamming window was used, taking samples every 20msecs.

The resonant frequencies produced in the vocal tract are referred to as formant frequencies or formants [27]. Although some studies in automatic recognition have looked at the first two formant frequencies (F1 and F2) [6] [26] [28], the formants have not been extensively researched. For this reason, in our experiments, the first five formant frequencies are extracted using Praat, which offers a formant tracking algorithm [29].

A. *Feature selection*

Based on the acoustic features described above and the literature relating to automatic emotion detection from speech, 133 prosodic features are calculated based, which are represented as contours: the pitch, the 12 MFCCs, the energy, and the first 5 formant frequencies. From these 19 contours, seven statistics have been extracted: the mean, the standard deviation, the minimum value, the maximum value, the range (max-min) of the original contour and the mean and standard deviation of the contour gradient. All the 133 measurements are shown in Table I (e.g., the mean of derivative for MFCC6 is the feature with ID 45).

TABLE I. THE OVERALL 133 SPEECH FEATURES. SHADED CELLS INDICATE THE SELECTED FEATURES

Prosodic Feature	*Mean*	*Std*	*Mean of derivative*	*Std of derivative*	*Max*	*Min*	*Range*
Pitch	1	2	3	4	5	6	7
MFCC1	8	9	10	11	12	13	14
MFCC2	15	16	17	18	19	20	21
MFCC3	22	23	24	25	26	27	28
MFCC4	29	30	31	32	33	34	35
MFCC5	36	37	38	39	40	41	42
MFCC6	43	44	45	46	47	48	49
MFCC7	50	51	52	53	54	55	56
MFCC8	57	58	59	60	61	62	63
MFCC9	64	65	66	67	68	69	70
MFCC10	71	72	73	74	75	76	77
MFCC11	78	79	80	81	82	83	84
MFCC12	85	86	87	88	89	90	91
Energy	92	93	94	95	96	97	98
F1	99	100	101	102	103	104	105
F2	106	107	108	109	110	111	112
F3	113	114	115	116	117	118	119
F4	120	121	122	123	124	125	126
F5	127	128	129	130	131	132	133

B. *Sound feature selection*

In order to select the most important prosodic features and optimise the classification time, the Bivariate Correlation procedure and especially the Spearman's rho correlation

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

coefficient was used. The Bivariate Correlations procedure computes Pearson's correlation coefficient, Spearman's rho, and Kendall's tau-b with their significance levels. Spearman's rho correlation coefficient is a measure of not linear association. Two symmetric quantitative variables or variables with ordered categories can be perfectly related by Spearman's rho correlation coefficient. If the data are not normally distributed or have ordered categories, the Kendall's tau-b or Spearman are chosen, which measure the association between rank orders. Correlation coefficients range in value from –1 (a perfect negative relationship) and +1 (a perfect positive relationship). A value of 0 indicates no linear relationship [4] [12].

For the method selection, the SPSS statistics software tool was used [30]. Among its features, the user may find modules for statistical data analysis, including descriptive statistics such as plots, frequencies, charts, and lists, as well as sophisticated inferential and multivariate statistical procedures like analysis of variance (ANOVA), factor analysis, cluster analysis, and categorical data analysis. Several correlation coefficients have been tested as shown in Table II in order to assess the feature selection combination that gives the optimum performance for our problem. The shaded cells in Table II shows the feature evaluator and search correlation coefficients that presented the best performance in the data set.

TABLE II. FEATURE EVALUATORS AND SEARCH METHODS THAT WERE CONSIDERED. SHADED CELLS INDICATE THE SELECTED COMBINATION

Feature evaluator (offered in SPSS)	***Feature search correlation coefficient (offered in SPSS)***
Bivariate Correlation	Pearson's
	Spearman's rho
	Kendall's tau-b

C. Spearman Correlation Coefficient

For each of the variables X and Y (Bivariate Correlation) separately, the observations are sorted into ascending order and replaced by their ranks. In situations where t observations are tied, the average rank is assigned. Each time t >1, the quantity t3-t is calculated and summed separately for each variable. These sums will be designated STx and STy. For each of the N observations, the difference between the rank of X and rank of Y is computed as seen in (2).

$$d_i = R(X_i) - R(Y_i) \quad (2)$$

Spearman's rho (p) [31] is calculated by (3), (4) and (5)

$$p_s = \frac{T_x + T_y - \sum_{i=1}^{N} d_i^2}{2\sqrt{T_x T_y}} \quad (3)$$

Where: $T_x = \frac{N^3 - N - ST_x}{12}$ (4)

and $T_y = \frac{N^3 - N - ST_y}{12}$ (5)

If Tx or Ty equals to 0, the statistic is not computed. The significance level is calculated assuming that, under the null hypothesis seen in (6), t is distributed as a t with N - 2 degrees of freedom. A one - or two-tailed significance level is printed depending on the user-selected option.

$$t = p_s \sqrt{\frac{N-2}{1-r_s^2}} \quad (6)$$

Two symmetric quantitative variables or variables with ordered categories can be perfectly related by Spearman's rho correlation coefficient. By this way we found the features that are highly correlated with the class variable (Emotion category). Spearman Correlation Coefficient (SCC) was selected as feature selection method, since it was proved after several tests, that it provides the most efficient sound measurements for the training of the Neural Network.

The combination of the above mentioned methods proposed 35 from the total of 133 features that were originally extracted. The shaded cells in Table I indicate the selected features. It can be seen, that 1 feature has been selected from pitch, namely the mean pitch. In addition, 29 features related to Mel Frequency Cepstral Coefficients were found important, while for the third prosodic group (energy) 5 features have been selected. Finally, none of the formant features have been found significantly important.

IV. CLASSIFICATION

In this research, the first classifier was an ANN, which was implemented following the multi-layer perceptron architecture, using WEKA software [24]. An artificial neural network (ANN), usually called "neural network" (NN), is a mathematical model or computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data [32]. After experimentation with various network topologies highest accuracy was found with 80 neurons in the hidden layer. The early stopping criterion was used based on a validation set consisting of 10% of the training set in the experiments and the number of training epochs was selected to be 400. This ensures that the training process stops when the mean-squared error (MSE) begins to increase on the validation set

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

avoiding the over-fitting problem in this problem. The learning and momentum rate were left to the default setting of WEKA (0.3 and 0.2 respectively). Error backpropagation was used as a training algorithm. Moreover, all neurons in WEKA follow the sigmoid activation function, while all attributes have been normalized for improved performance of the network.

In this research, MLP, PNN, and SVM classifiers were also used for classification and were performed by DTREG software [10]. Neural Networks are predictive models loosely based on the action of biological neurons. A multilayer perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable [33]. Although the implementation is very different, PNN are conceptually similar to K-Nearest Neighbor (k-NN) models. Probabilistic neural networks are forward feed networks built with three layers. They are derived from Bayes Decision Networks. They train quickly since the training is done in one pass of each training vector, rather than several. Probabilistic neural networks estimate the probability density function for each class based on the training samples. The probabilistic neural network uses Parzen or a similar probability density function. This is calculated for each test vector. This is what is used in the dot product against the input vector as described below. Usually a spherical Gaussian basis function is used, although many other functions work equally well [34].

In pattern recognition, the k-nearest neighbours algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: 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 neighbour [35]. The basic idea is that a predicted target value of an item is likely to be about the same as other items that have close values of the predictor variables.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training datapoints of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [36].

For PNN classifier with the 10 V-Fold validation method the highest accuracy was found with the Gaussian Kernel function and Sigma values for each variable and class. For the rest of the settings we used the default settings. The highest accuracy by SVM using the 10 V-Fold validation method was found with C-SVC SVM Model using RBF Kernel function.

A. Speaker dependent recognition in EMO-BD

The first experiment in our research corresponds to the problem of emotion recognition when the speakers are known to the classifier (i.e., at least on version of the each utterance/speaker is presented to the training set). Hence, this experiment corresponds to emotion recognition in a speaker dependent framework. For this case, from the 535 utterances provided in the Berlin Database we used four classifiers. First a MLP with 35-80-7 topology was trained using the 10-fold cross validation method. The training rate and the momentum were set 0.9 and 0.2 respectively, while the training stopped in 400 epochs. For the classification experiment the 10x10-fold stratified cross-validation method was employed over the data sets. Table III shows the confusion matrix for the 535 utterances of the Berlin database. Successful recognition is shown in the main diagonal. The overall success rate was 83.17%. However, considering the emotion classification in the two hyper-classes in Table IV, the correct classification reaches 95.1% for high arousal and 95.8% for low arousal emotions. Random Forest (RF) classifier was trained using the 10-fold cross validation method as well. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [37].

For the classification experiment the 10x10-fold stratified cross-validation method was employed over the data sets. Table V shows the confusion matrix for the 535 utterances of the Berlin database. Successful recognition is shown in the main diagonal. The overall success rate was 77,19%. However, considering the emotion classification in the two hyper-classes the correct classification reaches 94.3% for high arousal and 93% for low arousal emotions (as seen in Table VI).

After MLP, a PNN was trained. For the classification experiment the 10x10-fold stratified cross-validation method was employed over the data sets. Similarly as above, Table VII shows the confusion matrix for the 535 utterances of the

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

Berlin database. Successful recognition is shown in the main diagonal. The overall success rate was 94.1%. However, considering the emotion classification in the two hyper-classes the correct classification reaches 98.1% for high arousal and 97.7% for low arousal emotions (as seen in Table VIII).

Finally, a SVM was trained using the 10-fold cross validation method. Table IX shows the confusion matrix for the 535 utterances of the Berlin database. Successful recognition is shown in the main diagonal. The overall success rate was 84%. Again, considering the emotion classification, Table X depicts that in the two hyper-classes the correct classification reaches 97% for high arousal and 96.27% for low arousal emotions.

TABLE III. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR MLP CLASSIFIER

	High arousal emotions			Low arousal emotions			
	Anger	Happiness	Anxiety/fear	Boredom	Disgust	Sadness	Neutral
Anger	**115** (90.5%)	**6** (4.7%)	**5** (3.9%)	**0** (0.0%)	**1** (0.7%)	**0** (0.0%)	**0** (0.0%)
Happiness	**5** (7.0%)	**54** (76.0%)	**4** (5.6%)	**3** (4.2%)	**4** (5.6%)	**0** (0.0%)	**1** (1.4%)
Anxiety /fear	**8** (11.5%)	**5** (7.2%)	**52** (75.3%)	**0** (0.0%)	**2** (2.8%)	**1** (1.4%)	**1** (1.4%)
Boredom	**1** (1.2%)	**3** (3.7%)	**1** (1.2%)	**65** (80.2%)	**2** (2.4%)	**1** (1.2%)	**8** (9.8%)
Disgust	**0** (0.0%)	**3** (6.5%)	**3** (6.5%)	**5** (10.8%)	**34** (73.9%)	**1** (2.1%)	**0** (0.0%)
Sadness	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**5** (8.0%)	**0** (0.0%)	**53** (85.4%)	**4** (6.4%)
Neutral	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**5** (6.3%)	**0** (0.0%)	**2** (2.5%)	**72** (91.1%)

TABLE IV. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR MLP CLASSIFIER

	High arousal emotions	Low arousal emotions
High arousal emotions	(95.1%)	(4.9%)
Low arousal emotions	(4.2%)	(95.8%)

TABLE V. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR RANDOM FOREST CLASSIFIER

	High arousal emotions			Low arousal emotions			
	Anger	Happiness	Anxiety fear	Boredom	Disgust	Sadness	Neutral
Anger	**112** (88.2%)	**5** (3.9%)	**7** (5.5%)	**0** (0.0%)	**3** (2.4%)	**0** (0.0%)	**0** (0.0%)
Happiness	**9** (7.0%)	**51** (71.8%)	**4** (5.6%)	**1** (1.4%)	**6** (8.5%)	**0** (0.0%)	**0** (0.0%)
Anxiety /fear	**9** (13.0%)	**6** (8.7%)	**49** (71.0%)	**1** (1.4%)	**1** (1.4%)	**2** (2.8%)	**1** (1.4%)
Boredom	**0** (0.0%)	**2** (2.5%)	**3** (3.7%)	**55** (67.9%)	**2** (2.5%)	**4** (5.0%)	**15** (18.5%)
Disgust	**3** (6.5%)	**7** (15.2%)	**3** (6.5%)	**6** (13%)	**27** (58.7%)	**0** (0.0%)	**0** (0.0%)
Sadness	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**7** (11.3%)	**0** (0.0%)	**48** (77.4%)	**7** (11.3%)
Neutral	**1** (1.3%)	**0** (0.0%)	**0** (0.0%)	**9** (11.4%)	**0** (0.0%)	**2** (2.6%)	**67** (84.8%)

TABLE VI. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR RANDOM FOREST CLASSIFIER

	High arousal emotions	Low arousal emotions
High arousal emotions	(94.3%)	(5.7%)
Low arousal emotions	(7%)	(93%)

TABLE VII. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR PROBABILISTIC NEURAL NETWORK

	High arousal emotions			**Low arousal emotions**			
	Anger	Happiness	anxiety / fear	Boredom	Disgust	Sadness	Neutral
Anger	95.2% 121	0.8% 1	2.4% 3	0	0.8% 1	0	0.8% 1
Happiness	0	93% 66	2.8% 2	1.4% 1	2.8% 2	0	0
anxiety /fear	1.45% 1	1.45% 1	97.1% 67	0	0	0	0
Boredom	0	0	1.24% 1	91.35% 74	1.24% 1	0	6.17% 5
Disgust	0	2.18% 1	6.52% 3	0	91.3% 42	0	0
Sadness	0	0	0	0	0	98.4% 61	1.6% 1
neutral	0	0	1.3% 1	5% 4	0	1.3% 1	92.4% 73

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VIII. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR PROBABILISTIC NEURAL NETWORK

	High arousal emotions	Low arousal emotions
High arousal emotions	(98.1%)	(1.9%)
Low arousal emotions	(2.3%)	(97.7%)

TABLE IX. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR SVM

	High arousal emotions			**Low arousal emotions**			
	Anger	Happiness	anxiety / fear	Boredom	Disgust	Sadness	Neutral
Anger	88.9% 113	6.3% 8	4% 5	0	0.8% 1	0	0
Happiness	8.4% 6	83% 59	2.8% 2	0	5.8% 4	0	0
anxiety /fear	4.34% 3	5.8% 4	85.5% 59	1.45% 1	2,91% 2	0	0
Boredom	1.23% 1	0	1.23% 1	81.5% 66	4.93% 4	0	11.11% 9
Disgust	2.17% 1	8.7% 4	4.34% 2	10.8% 5	73.99% 34	0	0
Sadness	0	0	0	4.85% 3	0	88.7% 55	6.45% 4
neutral	1.37% 1	0	0	10.1% 8	0	2.53% 2	86% 68

TABLE X. CONFUSION MATRIX FOR SPEAKER-DEPENDENT EMOTION RECOGNITION FOR SVM

	High arousal emotions	Low arousal emotions
High arousal emotions	(97%)	(3%)
Low arousal emotions	(3.73%)	(96.27%)

B. *Speaker independent recognition in EMO-BD*

Speaker independent emotion recognition in Berlin database by MLP was evaluated averaging the results of five separate experiments. In each experiment, the measurements of a pair of speakers (e.g. , speaker 03 and speaker 08), were extracted from the training set and formed the testing set for the classifier. The pairs were selected in order to include one male and one female speaker at a time (Table XI-XVI).

The confusion matrix of Table XVII reveals that the MLP performance does not reach high accuracy. Overall, we are witnessing approximately 55% correct classification in the seven emotions. The 35-feature vector seems that it is not sufficient enough to distinguish the 7 emotions. However, observing the results in the two hyper-classes (low and high arousal), the recognition rate reach 89.1% for high arousal and 78.8% for low arousal emotions (see Table XVIII).

TABLE XI. TESTING AND TRAINING SETS FOR THE SPEAKER INDEPENDENT FRAMEWORK

Experiment no.	Testing set Utterances from speakers	Training set Utterances from speakers
1	10,11,12,15 (male), 09,13,14,16 (female)	03 (male), 08 (female)
2	03,11,12,15 (male), 08,13,14,16 (female)	10 (male), 09 (female)
3	03,10,12,15 (male), 08,09,14,16 (female)	11 (male), 13 (female)
4	03,10,11,15 (male), 08,09,13,16 (female)	12 (male), 14 (female)
5	03,10,11,12 (male), 08,09,13,14 (female)	15 (male), 16 (female)

TABLE XII. EXPERIMENT 1: EVALUATION IN SPEAKERS 03 AND 08.

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	**5** (19.2%)	**20** (76.9%)	**1** (3.8%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)
Happiness	**0** (0.0%)	**15** (83.3%)	**0** (0.0%)	**0** (0.0%)	**3** (16.7%)	**0** (0.0%)	**0** (0.0%)
anxiety /fear	**0** (0.0%)	**5** (50.0%)	**1** (10.0%)	**4** (40.0%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)
Boredom	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**12** (80.0%)	**3** (20.0%)	**0** (0.0%)	**0** (0.0%)
Disgust	**0** (0.0%)	**1** (100.0%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)
Sadness	**0** (0.0%)	**2** (12.5%)	**0** (0.0%)	**7** (43.8%)	**2** (12.5%)	**5** (31.3%)	**0** (0.0%)
neutral	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)	**11** (52.4%)	**0** (0.0%)	**0** (0.0%)	**10** (47.6%)

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE XIII. EXPERIMENT 2: EVALUATION IN SPEAKERS 10 AND 09

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	18 (78.3%)	1 (4.3%)	2 (8.7%)	0 (0.0%)	2 (8.7%)	0 (0.0%)	0 (0.0%)
Happiness	2 (25.0%)	2 (25.0%)	3 (37.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
anxiety /fear	0 (0.0%)	0 (0.0%)	8 (88.9%)	1 (11.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Boredom	0 (0.0%)	2 (16.7%)	0 (0.0%)	10 (83.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Disgust	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (77.8%)	2 (22.2%)	0 (0.0%)	0 (0.0%)
Sadness	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (28.6%)	0 (0.0%)	3 (42.9%)	2 (28.6%)
neutral	1 (7.7%)	0 (0.0%)	0 (0.0%)	6 (46.2%)	0 (0.0%)	2 (15.4%)	4 (30.8%)

TABLE XIV. EXPERIMENT 3: EVALUATION IN SPEAKERS 11 AND 13.

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	14 (63.6%)	0 (0.0%)	1 (4.5%)	0 (0.0%)	6 (27.3%)	0 (0.0%)	1 (4.5%)
Happiness	6 (33.3%)	2 (11.1%)	2 (11.1%)	0 (0.0%)	6 (33.3%)	0 (0.0%)	2 (11.1%)
anxiety /fear	13 (76.5%)	0 (0.0%)	1 (5.9%)	1 (5.9%)	1 (5.9%)	1 (5.9%)	0 (0.0%)
Boredom	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (38.9%)	0 (0.0%)	1 (5.6%)	10 (55.6%)
Disgust	2 (20.0%)	0 (0.0%)	0 (0.0%)	1 (10.0%)	7 (70.0%)	0 (0.0%)	0 (0.0%)
Sadness	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (75.0%)	3 (25.0%)
neutral	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (5.6%)	0 (0.0%)	1 (5.6%)	16 (88.9%)

TABLE XV. EXPERIMENT 4: EVALUATION IN SPEAKERS 12 AND 14

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	14 (50.0%)	2 (7.1%)	12 (42.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Happiness	5 (50.0%)	4 (40.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (10.0%)
anxiety /fear	8 (44.4%)	1 (5.6%)	9 (50.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Boredom	1 (7.7%)	0 (0.0%)	0 (0.0%)	5 (38.5%)	0 (0.0%)	6 (46.2%)	1 (7.7%)
Disgust	3 (30.0%)	0 (0.0%)	0 (0.0%)	1 (10.0%)	4 (40.0%)	0 (0.0%)	2 (20.0%)
Sadness	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (14.3%)	0 (0.0%)	10 (71.4%)	2 (14.3%)
neutral	0 (0.0%)	2 (18.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (54.5%)	3 (27.3%)

TABLE XVI. EXPERIMENT 5: EVALUATION IN SPEAKERS 15 AND 16

	High arousal emotions			Low arousal emotions		
	anger	happiness	anxiety/ fear	boredom	disgust	sadness
Anger	25 (92.6%)	0 (0.0%)	2 (7.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Happiness	9 (52.9%)	7 (41.2%)	1 (5.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
anxiety /fear	7 (46.7%)	0 (0.0%)	8 (53.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Boredom	0 (0.0%)	0 (0.0%)	12 (52.2%)	9 (39.1%)	0 (0.0%)	2 (8.7%)
Disgust	1 (6.3%)	8 (50.0%)	2 (12.5%)	2 (12.5%)	3 (18.8%)	0 (0.0%)
Sadness	0 (0.0%)	0 (0.0%)	1 (7.7%)	0 (0.0%)	1 (7.7%)	10 (76.9%)
neutral	0 (0.0%)	0 (0.0%)	3 (17.6%)	2 (11.8%)	0 (0.0%)	1 (5.9%)

TABLE XVII. MLP CLASSIFIER- OVERALL PERFORMANCE IN 7 EMOTION CLASSES AFTER THE 5 SPEAKER-INDEPENDENT EXPERIMENTS

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	83 (65.8%)	22 (17.4%)	18 (14.2%)	0(0.0%)	1 (0.7%)	2(1.5%)	1(0.7%)
Happiness	19 (26.7%)	27(38.0%)	9 (12.6%)	3 (4.2%)	11 (15.4%)	0(0.0%)	2(2.8%)
anxiety /fear	23 (33.3%)	4(5.7%)	33 (47.8%)	0(0.0%)	5(7.2%)	1(1.4%)	3(4.3%)
Boredom	1(1.2%)	1 (1.2%)	15 (18.5%)	48(59.2%)	1 (1.2%)	5(6.1%)	10 (12.3%)
Disgust	3(6.5%)	16 (34.7%)	4(8.6%)	3 (6.5%)	16 (34.7%)	2(4.3%)	2(4.3%)
Sadness	0(0.0%)	1(1.6%)	1 (1.6%)	8(12.9%)	2 (3.2%)	39(62.9%)	11 (17.7%)
neutral	1(1.2%)	10 (12.5%)	4(5.0%)	10(12.5%)	0 (0.0%)	5 (6.2%)	50 (62.5%)

TABLE XVIII. MLP CLASSIFIER

	High arousal emotions	Low arousal emotions
High arousal emotions	**238** (89.1%)	**29** (10.8%)
Low arousal emotions	**57** (21.1%)	**212** (78.8%)

The confusion matrices of Table XIX show that the RF performance reaches lower accuracy than MLP classifier (approximately 49% correct classification in the seven emotions). However, the results is higher for RF in the two hyper-classes (low and high arousal), where the recognition rate reaches 89.4% for high arousal and 82.52% for low arousal emotions (see Table XX).

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE XIX. RANDOM FOREST CLASSIFIER

	High arousal emotions			Low arousal emotions			
	Anger	Happiness	Anxiety/ fear	Boredom	Disgust	Sadness	Neutral
Anger	77 (73.33%)	12 (11.42%)	6 (5.7%)	1 (0.95%)	1 (0.95%)	5 (4.76%)	3 (2.8%)
Happiness	24(33.8%)	23 (32.4%)	18 (25.35%)	0 (0%)	6 (8.45%)	0 (0.0%)	0 (0.0%)
Anxiety /fear	15 (21.7%)	14 (20.2%)	30 (43.4%)	2 (2.9%)	6 (8.7%)	1 (1.4%)	1 (1.4%)
Boredom	1 (1.23%)	2 (2.46%)	12 (14.8%)	33 (40.7%)	7 (8.64%)	9 (11.11%)	17 (20.9%)
Disgust	4 (8.5%)	10 (21.27%)	10 (21.27%)	9 (19.14%)	9(19.14%)	3 (6.38%)	2 (4.3%)
Sadness	0 (0.0%)	0 (0.0%)	3(4.8%)	11 (17.74%)	4 (6.45%)	32(51.61%)	12 (19.35%)
Neutral	1 (11.11%)	2 (2.5%)	2 (2.5%)	33 (41.78%)	4 (5%)	8 (10.13%)	29 (36.7%)

TABLE XX. RANDOM FOREST

	High arousal emotions	Low arousal emotions
High arousal emotions	**219** (89.4%)	**26** (10.6%)
Low arousal emotions	**47** (17.48%)	**222** (82.52%)

TABLE XXI. SUPPORT VECTOR MACHINE - OVERALL PERFORMANCE

	High arousal emotions			Low arousal emotions			
	anger	happiness	anxiety/ fear	boredom	disgust	sadness	neutral
Anger	84.38%	6.25%	**9.37%**	0.0%	0.0 %	0.0%	0.0 %
Happiness	**5.55 %**	**88.9%**	**5.55%**	0.0%	0.0 %	0.0%	0.0 %
anxiety /fear	**5.6 %**	0.0%	**94.4%**	0.0%	0.0 %	0.0%	0.0 %
Boredom	**5%**	**5%**	**10%**	**55%**	**15 %**	0.0%	**10 %**
Disgust	0.0 %	**27.27 %**	0.0%	**18.18%**	**54.55%**	**0** (0.0%)	0.0 %
Sadness	0.0 %	0.0%	0.0%	0.0%	0.0 %	**80%**	**20 %**
neutral	**5%**	0.0%	0.0%	**10%**	**0** (0.0%)	**5%**	**80 %**

TABLE XXII. SUPPORT VECTOR MACHINE

	High arousal emotions	Low arousal emotions
High arousal emotions	(89.1%)	(10.9%)
Low arousal emotions	(21.2%)	(78.8%)

Similarly, the confusion matrices of Table XXI and Table XXII reveals that the SVMs performance (approximately 78% correct classification in the seven emotions) reach higher accuracy than MLP (approximately 53% correct classification in the seven emotions). However, observing the results for MLP in the two hyper-classes (low and high arousal), the recognition rate reach 89.1% for high arousal and 78.8% for low arousal emotions, while the results is surprising higher for SVM in the two hyper-classes (low and high arousal), the recognition rate reaches 100% for high arousal and 87% for low arousal emotions. We did not use the PNN classifier for Independent Emotion Recognition because the DTREG Tool had problems during the classification procedures

V. CONCLUSION AND DISCUSSION

The literature in speech emotion detection is not very rich and researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together.

The researchers usually deal with elicited and acted emotions in a lab setting from few actors, just like in our case. However, in the real problem, different individuals reveal their emotions in a diverse degree and manner. There are also many differences between acted and spontaneous speech. Speaker-independent detection of negative emotional states from acted and real-world speech, was investigated in [21]. The experimentations demonstrated some important differences on recognizing acted versus non-acted speech, which cause significant drop of performance, for the real-world data.

Although it is impossible to accurately compare recognition accuracies from this study to other due to different data sets used, the feature set implemented in this work seems to be promising for further research.

Concluding this paper, the 35-input vector, seems to be quite promising for speaker independent recognition in terms of high and low arousal emotions when tested in Berlin database. Nevertheless, this vector is not sufficient enough to describe the intra class variations of the two hyper classes.

The major finding in this work is that PNN classifier achieved almost perfect classification (94%) in speaker dependent emotion recognition. This finding suggests that PNN proved to be the most adequate classifier for the dependent emotion recognition field. As well, in speaker independent emotion recognition, SVMs overall success rate was very high (78%), and the surprising finding is that SVM achieved perfect correct classification in High arousal emotions(100%), and significant success rate in Low arousal emotions(87%).

A future work should encompass an Hybrid Mixed Classification Model combining PNN, SVM and maybe

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

more classifiers so as to recognize emotions in non-acted speech.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction”, Proceedings of the IEEE, vol. 91, Issue 9, Sept. 2003, pp. 1370–1390, doi: 10.1109/JPROC. 2003.817122.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction.”, IEEE Signal Processing Magazine, vol. 18, Issue 1, Jan. 2001, pp. pp. 32–80, doi: 10.1109/79.911197.

[3] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human–computer dialog”, Proc. International Conference on Spoken Language Processing (ICSLP 2002), ISCA, Dec. 2002, pp. 2037-2040.

[4] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos, "Statistical Evaluation of Speech Features for Emotion Recognition," icdt, pp.121-126, 2009 Fourth International Conference on Digital Telecommunications, 2009.

[5] L. C. De Silva and P. C. Ng, “Bimodal emotion recognition”, Proc. 4th IEEE Intrnational Conference on Automatic Face and Gesture Recognition (FG), IEEE, March 2000, pp. 332–335, doi: 10.1109/AFGR.2000.840655.

[6] V. Petrushin, “Emotion recognition in speech signal: experimental study, development, and application”, Proc. Sixth International Conference on Spoken Language Processing (ICSLP 2000), ISCA, Oct.2000. vol 2, pp. 222-225.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of german emotional speech” Proc. 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP 2005), ISCA, Sept. 2005, pp. 1517-1520.

[8] P. J. Lang, “The Emotion Probe: Studies of Motivation and Attention”, American Psychologist, vol 50, no. 5, May 1995, pp. 372-85.

[9] S. Kim, P. Georgiou, S. Lee, and S. Narayanan. “Real-time emotion detection system using speech: Multi-modal fusion of different timescale features”, Proc. 9th IEEE Workshop on Multimedia Signal Processing Workshop, Oct. 2007, pp. 48-51, doi 10.1109/MMSP.2007.4412815.

[10] DTREG, [Computer program], http://www.dtreg.com/; last access date [June 10, 2010].

[11] Jef Raskin: The humane interface. New directions for designing interactive systems. Addison-Wesley, Boston 2000 ISBN 0-201-37937-6

[12] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos, " Comparison of Different Classifiers for Emotion Recognition," pci, pp.121-126, 2009 Fourth PanHellenic Conference on Informatics, 2009.

[13] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, “Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face”, Proc. 9th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN), IEEE, Sept. 2000, pp. 178–183, doi: 10.1109/ROMAN.2000.892491.

[14] V. Hozjan and Z. Kacic, “Context-independent multilingual emotion recognition from speech signals”, International journal of Speech Technology, vol. 6, July 2003, pp. 311–320, doi: 10.1023/A:1023426522496.

[15] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres”, Speech Communication, vol. 49, issue 2, Feb. 2007, pp. 98–112, doi: 10.1016/j.specom.2006.11.004.

[16] Fant, G. (1960). Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands.

[17] T.Bänziger and K.R.Scherer, “The role of intonation in emotional expression”, Speech Communication, vol.46, issues 3-4, July 2005, pp. 252-267, doi:. 10.1016/j.specom.2005.02.016.

[18] W. H. Abdulla and N. K. Kasabov, “Improving speech recognition performance through gender separation”, Proc. Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES), University of Otago Printery, Nov. 2001, pages 218–222.

[19] Y. Wang and L. Guan, “Recognizing human emotion from audiovisual information”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASP), March 2005, pp. 1125-1128.

[20] T.Vogt and E. Andre, “Improving Automatic Emotion Recognition from Speech via Gender Differentiation”, Proc. Language Resources and Evaluation Conference (LREC), May 2006, pp. 1123–1126.

[21] T. P. Kostoulas and N. Fakotakis, “A Speaker Dependent Emotion Recognition Framework”, Proc. 5th International Symposium, Communication Systems, Networks and Digital Signal Processing (CSNDSP), University of Patras, July 2006, pp. 305-309.

[22] M.Fingerhut, "Music Information Retrieval, or how to search for (and maybe find) music and do away with incipits", Proc. joint Conf. International Association of Music Libraries, International Association of Sound and Audiovisual Archives (IAML-IASA), Aug. 2004, CD proceedings.

[23] D. Talkin, “A Robust Algorithm for Pitch Tracking (RAPT)”, in Speech Coding & Synthesis, W. B. Kleijn, K. K. Paliwal, Eds., Elsevier Science Inc., 1995.

[24] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html; last access date [June 10, 2010].

[25] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/; last access date [June 10, 2010].

[26] K.R. Scherer, “Vocal communication of emotion: a review of research paradigms”, Speech Communication, vol 40, issues 1-2, April 2003, pp. 227–256, doi: 10.1016/S0167-6393(02)00084-5.

[27] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[28] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, “Emotion recognition based on phoneme classes”, Proc. 8th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP 2004), Oct. 2004, pp. 889-892.

[29] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (Version 4.6.09)”, 2005 [Computer program], http://www.praat.org/; last access date [June 10, 2010].

[30] http://www.spss.com/corpinfo/?source=homepage&hpzone=nav ; last access date [June 10, 2010].

[31] S. Siegel, Non-parametric Statistics for the Behavioral Sciences. McGraw-Hill Book Co., Inc., New York, 1956.

[32] Neural Networks as Cybernetic Systems 2nd and revised edition, Holk Cruse.

[33] Haykin, Simon (1998). Neural Networks: A Comprehensive Foundation (2 ed.). Prentice Hall. ISBN 0132733501.

[34] D. Specht, Probabilistic neural networks Neural Networks vol. 3 pp.109-118,1990.

[35] Belur V. Dasarathy, ed (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. ISBN 0-8186-8930-7.

[36] Corinna Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, 20, 1995.

[37] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.

[38] T. Kostoulas, T. Ganchev, and N. Fakotakis, “Study on speaker-independent emotion recognition from speech on real-world data”, in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, A. Esposito, Eds. Springer Verlag, Berlin, Heidelberg, 2008, pp. 235-242.

2010, © Copyright by authors, Published under agreement with IARIA - www.iaria.org