Expert Systems With Applications 42 (2015) 9554–9564

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

New approach in quantification of emotional intensity from the speech signal: emotional temperature

Jesús B. Alonso*, Josué Cabrera, Manuel Medina, Carlos M. Travieso

Instituto Universitario para el Desarrollo Tecnológico y la Innovación en Comunicaciones (IDeTIC), Universitdad de Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Keywords:
Emotional speech recognition
Pattern recognition
Emotional intensity
Emotional temperature

ABSTRACT

The automatic speech emotion recognition has a huge potential in applications of fields such as psychology, psychiatry and the affective computing technology. The spontaneous speech is continuous, where the emotions are expressed in certain moments of the dialogue, given emotional turns. Therefore, it is necessary that the real-time applications are capable of detecting changes in the speaker's affective state. In this paper, we emphasize on recognizing activation from speech using a few feature set obtained from a temporal segmentation of the speech signal of different language like German, English and Polish. The feature set includes two prosodic features and four paralinguistic features related to the pitch and spectral energy balance. This segmentation and feature set are suitable for real-time emotion applications because they allow detect changes in the emotional state with very low processing times. The German Corpus EMO-DB (Berlin Database of Emotional Speech), the English Corpus LDC (Emotional Prosody Speech and Transcripts database) and the Polish Emotional Speech Database are used to train the Support Vector Machine (SVM) classifier and for gender-dependent activation recognition. The results are analyzed for each speech emotion with gender-dependent separately and obtained accuracies of 94.9%, 88.32% and 90% for EMO-DB, LDC and Polish databases respectively. This new approach provides a comparable performance with lower complexity than other approaches for real-time applications, thus making it an appealing alternative, may assist in the future development of automatic speech emotion recognition systems with continuous tracking.

1. Introduction

The study of emotions is a challenge (Scherer, 1981) which recently we have become aware of their potential (Picard, 1999). Affective expression have become a growing field of research in psychology (James, 1884; Russell, 1997), psychiatry (Scherer, 1981), linguistic (Kitayama & Markus, 1994), phonetic (Roach, 2000) and computer science (Cowie et al., 2001; Oudeyer, 2002; Roy & Pentland, 1996; Petrushin, 2000).

Currently, the researches in emotion recognition are focused mainly on the online sentiment analysis in social media (Desmet & Hoste, 2013), online news (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014) or blogs (Li & Xu, 2014) using the text mining, and on the Human Computer Interaction (HCI) using facial emotion recognition (Ali, Hariharan, Yaacob, & Adom, 2015; Zhang, Zhang, & Hossain, 2015), emotion generation (Alepis & Virvou, 2011; Ammar, Neji, Alimi, & Gouardères, 2010; López-Ludeña, Barra-Chicote, Lutfi, Montero, & San-Segundo, 2013) and Speech Emotion Recognition (SER).

The automatic SER systems identify the emotional state of human from the speech signal. In the state-of-art many researches are focused on the use of SER in applications of HCI for the real-world. One potential application is the detection of the emotional state in telephone call-center conversations, providing feedback to an operator or supervisor for monitoring purposes, or enter a specific mode designed to resolve problems (Neiberg & Elenius, 2008; Petrushin, 1999; Petrushin, 2000; Yacoub, Simske, Lin, & Burns, 2003). Another application is sorting voice mail messages according to the emotions expressed by the caller in a medical emergency call-center (Vidrascu & Devillers, 2007). In the field of diagnostic and therapeutic tools, the SER systems can give support to traditional methods for treatment of mental disorders trough game-oriented interfaces with spoken interaction (Kostoulas et al., 2012); detect mental illness using speech as a predictor of depression, suicidality and mood transitions (Cummins et al., 2015; Karam et al., 2014); and may assistant in early diagnosis of patients with Parkinson's disease (Zhao, Rudzicz, Carvalho, Márquez-Chin, & Livingstone, 2014) and Alzheimer's disease (Lopez-de-Ipiña et al., 2013) which affect a patient's ability to produce an emotional tone of voice. Others applications of SER include safety in automotive

* Corresponding author. Tel.: +34 928452863. Fax: +34928451243.
E-mail addresses: jalonso@dsc.ulpg.es (J.B. Alonso), jcabrera@idetic.eu (J. Cabrera), manuel.medina@ulpgc.es
(M. Medina), ctravieso@dsc.ulpgc.es (C.M. Travieso).

http://dx.doi.org/10.1016/j.eswa.2015.07.062
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

(Nass et al., 2005; Tawari & Trivedi, 2010), security systems (Cabrera et al., 2015), help systems to autistic people (Petrushin, 2000), intelligent toys, lie detection, learning environment, educational software, virtual agents, entertainment and games (Cowie et al., 2001). The proposed approach is focused to be suitable for the development of this type of applications that are required a continuous tracking of the speech signal, a low complexity and low processing time.

There is no agreement on the number of emotions to be analyzed either, yet it is widely accepted model emotions as a set of categories or independent discrete types in which there are a few basic or primitive emotions (Cowie & Cornelius, 2003) and the rest of them, secondary emotions, resulting from a combination of the first. Most researches focus on four basic emotions: anger, fear, sadness and happy (Chavhan, Yelure, & Tayade, 2015; Julia & Iftekharuddin, 2005; Zhao et al., 2014); or in six emotions: happy, anger, fear, sadness, surprise and disgust (Balti & Elmaghraby, 2014; Kanagaraj, Shahina, Devosh, & Kamalakannan, 2014; Ooi, Seng, Ang, & Chew, 2014; Rabiei & Gasparetto, 2014; Razak, Komiya, Izani, & Abidin, 2005). Some authors extend the list distinguishing between cold and hot anger (Yacoub et al., 2003), or adding other emotions such as boredom (Le & Lee, 2014; Tawari & Trivedi, 2010; Xiao, Dellandrea, Dou, & Chen, 2007), pride, panic (Petrushin, 1999) or love. The "neutral" emotion is used like an intermediate state when switching between two different emotions.

Nevertheless, other researches, focused on the development of real-time applications and on the detection of changes in the speaker's affective state, emphasize the usefulness of represent emotions in an evaluation plane in terms of two or more levels or continuous dimensions (Laukka, 2004; Scherer, 1981) that allow recognize more easily gradual emotional transitions and changes in intensity, which is not possible if trying to recognize emotions directly (Coutinho, Deng, & Schuller, 2014; Goudbeek & Scherer, 2010; Harimi, Shahzadi, & Ahmadyfard, 2014; Lika, Seldon, & Kiong, 2014; Mencattini et al., 2014; Pohjalainen & Alku, 2014; Poon-Feng, Huang, Dong, & Li, 2014; Wöllmer et al., 2008; Wöllmer, Schuller, Eyben, & Rigoll, 2010). In practice, activation and valence level are the two most commonly used dimensions (Fig. 1). The activation is related to the intensity perceived from the emotion: if we take into consideration that when in a certain emotional state there is a certain tendency to behave in a particular way, the activation would be the measure to the intensity of that stimulus. The valence level, on the other hand, has to do with the pleasant sensation perceived from that stimulus, for example 'positive' in the case of happy or 'negative' in the case of sadness. This study is focused on activation recognition. To represent the scale of activation we have considered five emotions: anger, happy, neutral, boredom and sadness. These emotions are the top most emotion recorded and analyzed by the researchers in the speech emotion field according to on a review of 32 emotional speech databases (Ververidis & Kotropoulos, 2003). The emotions anger and happy represent high activation, and the emotions neutral state, boredom and sadness represent low activation.

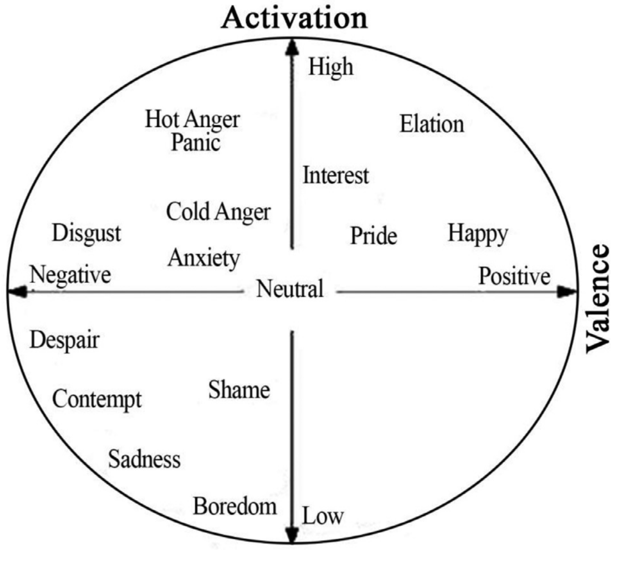

Fig. 1. Activation and valence space.

Emotional corpora may be categorized into three classes: acted, induced and natural. With the aim to develop applications for the real world it is suitable to use natural emotional speech, acquired from real-life situation where the emotions are authentic and not manipulated in any way, or induced speech where the subjects do certain task to induce the intended emotion. However, the natural speech has some legal and ethical issues and creates challenge to the researchers to individually classify to the correct emotion. On the other hand, in the induced speech the subjects may fake the reaction, resulting in a misleading emotional speech feedback (Kamaruddin, Wahab, & Quek, 2012). For these reasons most researches are focused to use acted emotion corpora recorder by actors in studio conditions where many aspects of the recording can be carefully and systematically controlled (Chenchah & Lachiri, 2014). In the state-of-the-art there is a wide variety of databases of acted speech using several languages and different emotions (El Ayadi, Kamel, & Karray, 2011). It is noticed that human are able to recognize emotion cross-culturally due to universality of the emotional acoustic features in the speech. In consequence, some researchers have used several databases of different languages merged while others researches have opted to maintain the unit of the language and other aspect such as gender (Chavhan et al., 2015; Iriya & Arjona Ramirez, 2014; Mencattini et al., 2014). In this work, as first approach we define text-independent, speaker-independent, gender-dependent and language–dependent study. In this sense, we have used the Berlin Database of Emotional Speech EMO-DB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), the LDC Emotional Prosody Speech and Transcripts database (Liberman, Davis, Grossman, Martey, & Bell, 2002) and the Polish Emotional Speech Database (Staroniewicz & Majewski, 2009) separately.

In general, the speech emotion analysis in previous research assume that each analyzed utterance contains only one emotional state within the given recorded interval (Kwon, Chan, Hao, & Lee, 2003; Lee & Narayanan, 2003). Perhaps this statement is questionable recordings of certain duration. Human emotion is a continuum and an automatic emotion recognition system must be able to recognise it as such (Wöllmer et al., 2008). In real-time applications there are problems when the emotion recognition are applied in continuous speech since the emotions are expressed explicitly only in certain moments of the dialogue. Single pitch or energy values are not meaningful for emotions, but rather their behavior over time. In this case, a temporal segmentation and find information of each emotionally salient segment is more appropriate (Balti & Elmaghraby, 2014; Fan, Xu, Wu, & Cai, 2014; Pao, Chien, Yeh, Chen, & Cheng, 2007; Vogt & André, 2005; Wöllmer et al., 2008) because emotion changes can occur very quickly and the segment length sets the temporal resolution of recognizable changes. In this work, we have used temporal segmentation with a fixed segment length. Then, a feature set is extracted of each segment and emotion is recognized individually. In this way, this approach can support emotion recognition from continuous emotional speech, finding out the changing points.

The previous researches in automatic speech emotion recognition presents features extraction based on the characterization of prosodic aspects, such as pitch contour, energy contour, the durations of phonations (Goudbeek & Scherer, 2010; Kwon et al., 2003; Lee & Narayanan, 2003; Mustafa & Ainon, 2013; Ooi et al., 2014; Petrushin, 1999; Sankar, 1988; Vogt & André, 2005) or Teager Energy Operator

(TEO) (Harimi et al., 2014; Mencattini et al., 2014; Ooi et al., 2014), aspects related to the way that a speaker communicates the grammatical structure and the lexical stress; and on the characterization of paralinguistic aspects, such as the first formant (Kwon et al., 2003; Lee & Narayanan, 2003; Pao et al., 2007; Petrushin, 1999; Rabiei & Gasparetto, 2014), the concentration of energy in different frequency bands (Altun & Polat, 2009; Poon-Feng et al., 2014; Scherer, 1987; Schirmer, Striano, & Friederici, 2005; Wöllmer et al., 2010), the Linear Prediction Coefficients (LPC) (Altun & Polat, 2009; Pao et al., 2007; Razak et al., 2005), Linear Prediction Cepstral Coefficients (LPCC) (Pao et al., 2007), Log Frequency Power Coefficients (LFPC) (Iriya & Arjona Ramirez, 2014) or Cepstral Mel-Frequency Components (MFCC) (Fan et al., 2014; Julia & Iftekharuddin, 2005; Kwon et al., 2003; Ooi et al., 2014; Pao et al., 2007; Vogt & André, 2005; Wöllmer et al., 2010), among others, are related to the manifestations of the emotional state in the phonatory system due to the autonomic nervous system and therefore cannot be controlled by the speaker. The classical approach in speech emotion recognition uses wide feature set, between 13–1941 features, of prosodic and paralinguistic features (Altun & Polat, 2009; Neiberg & Elenius, 2008; Ooi et al., 2014; Poon-Feng et al., 2014; Vidrascu & Devillers, 2007). This approach is not suitable for applications in real-time or for control of changes of speaker's emotional state due to the complexity of some of the features and the time processing time that carries such a wide feature set. In this work, we proposed reduced and robust feature set that includes two prosodic features and four paralinguistic features related to the pitch and spectral energy balance. The low computation of this feature set allows the fulfilment to the performance that time-real applications and automatic emotion variation detection systems in continuous speech should ensure.

These features may be used directly in models capable of classifying sequences of data. In the state-of-art, previous researches have used mainly classifiers such as Gaussian Mixture Models (GMM) (Iriya & Arjona Ramirez, 2014), Hidden Markov Models (HMM) (Chenchah & Lachiri, 2014), Neural Network (NN) (Petrushin, 1999), Support Vector Machines (SVM) (Amol & Guddeti, 2014; Burges, 1998; Chavhan et al., 2015; Julia & Iftekharuddin, 2005; Tawari & Trivedi, 2010; Vidrascu & Devillers, 2007) and RBF Neural Network (NN) (Ooi et al., 2014), among others, in addition to modified models (Harimi et al., 2014; Kwon et al., 2003; Pao et al., 2007) and combined or hierarchical classification systems (Kanagaraj et al., 2014; Le & Lee, 2014; Pao et al., 2007; Xiao et al., 2007) of these classifiers. We proposed a framework that exploits the classifier SVM for classifying emotionally segments extracted from temporal segmentation of the continuous speech. We introduce a hierarchical approach in two stages that generate a quantification of activation from the speech signal which we have called *Emotional Temperature*. The recognition stage is simple to implement with less computational efforts than other approaches of hierarchical classification systems.

This paper focuses on the complex task of quantified activation using a method with low processing time and capable of detecting changes in the speaker's affective state to its future use in real-time applications. The proposed method uses temporal segmentation and to characterize each segment by means of a reduced and robust feature set. Then, each emotional segment is classified through a hierarchical classification system whose resulting measure, called *Emotional Temperature*, quantifies the activation level of the emotional speech signal. Using this measure is even possible to do an automatically discrimination between high and low activation. The calculation of this measure has a low computational cost which is ideal for real-time applications. Section 2 describes the material and methods, the databases used in the study are presented and the measure *Emotional Temperature* is explained. Section 3 shows the results obtained and the discussion. Finally, Section 4 is devoted to conclusions.

2. Material and methods

This section describes the different steps of the experimental procedure: pre-processing, feature extraction, classifier, the measure *Emotional Temperature* and the evaluation of the proposed strategy. In the evaluation step, the proposed feature set and *Emotional Temperature* are evaluated using three databases in different languages in order to validate their usefulness in discriminating between high activation and low activation.

2.1. Databases

This section describes the public databases of emotional speech used in this paper: the Berlin emotional speech database (Burkhardt et al., 2005), the Emotional Prosody Speech and Transcripts of the Linguistic Data Consortium (LDC) (Liberman et al., 2002) and the Polish Emotional Speech Database (Staroniewicz & Majewski, 2009). The databases include discrete emotions produced by actors. The frequency sample at which the data were recorded is different for each database: 16 kHz in the Berlin emotional database, 22 kHz in the LDC database and 44.1 kHz in the Polish emotional database. The databases are resampled at 16 kHz in our experiments.

2.1.1. Berlin emotional speech database

The Berlin emotional speech database (EMO-DB) was produced by ten actors, 5 males and 5 females. Each actor uttered ten sentences in German, 5 short sentences (1.5 s approximately) and 5 longer sentences (4 s approximately). Actors simulated seven emotions: happiness, fear, anger, sadness, boredom, disgust and neutral state. The database only includes the utterances scoring higher than 80% emotion recognition rate in a subjective listening test.

2.1.2. LDC emotional prosody speech and transcripts database

The LDC database contains emotional speech in English language simulated from professional actors. The utterances are short phrases (around 1 s of duration) consisting of dates and numbers. The entire database consists of 7 actors expressing 15 emotions: anxiety, boredom, cold anger, contempt, despair, disgust, elation, happy, hot anger, interest, neutral state, panic, pride, sadness and shame. In the transcription of the database, neutral utterances were labeled as neutral and neutral with distance information (tête à tête, conversation and distant). We only consider the utterances labeled as neutral (without distance information).

2.1.3. Polish emotional speech database

The Polish Emotional Speech Database comprises 240 recordings from 8 actors (4 females and 4 males). Recordings for every speaker were made during a single session. Each speaker utters five different sentences in Polish language with six kinds of emotional load: joy, boredom, fear, anger, sadness and neutral. Each emotional state comprises 40 recordings.

2.2. Preprocessing

Each recording from database $\{s(n)\}$ is preprocessed using a Voice Activity Detector (VAD) (Boll, 1979) in order to remove the silences from the speech samples of the database. From each speech signal the DC component is removed and the z-score normalization is made. After that, the speech signal $\{s(n)\}$ is windowed by a hamming window of 0.5 s overlapped 50% (Pao et al., 2007) is called$\{w(n)\}$.

2.3. Feature extraction

Two prosodic features and four paralinguistic features related to the pitch and spectral energy, respectively, are estimated from each voiced frame. These features are chosen for several reasons: firstly,

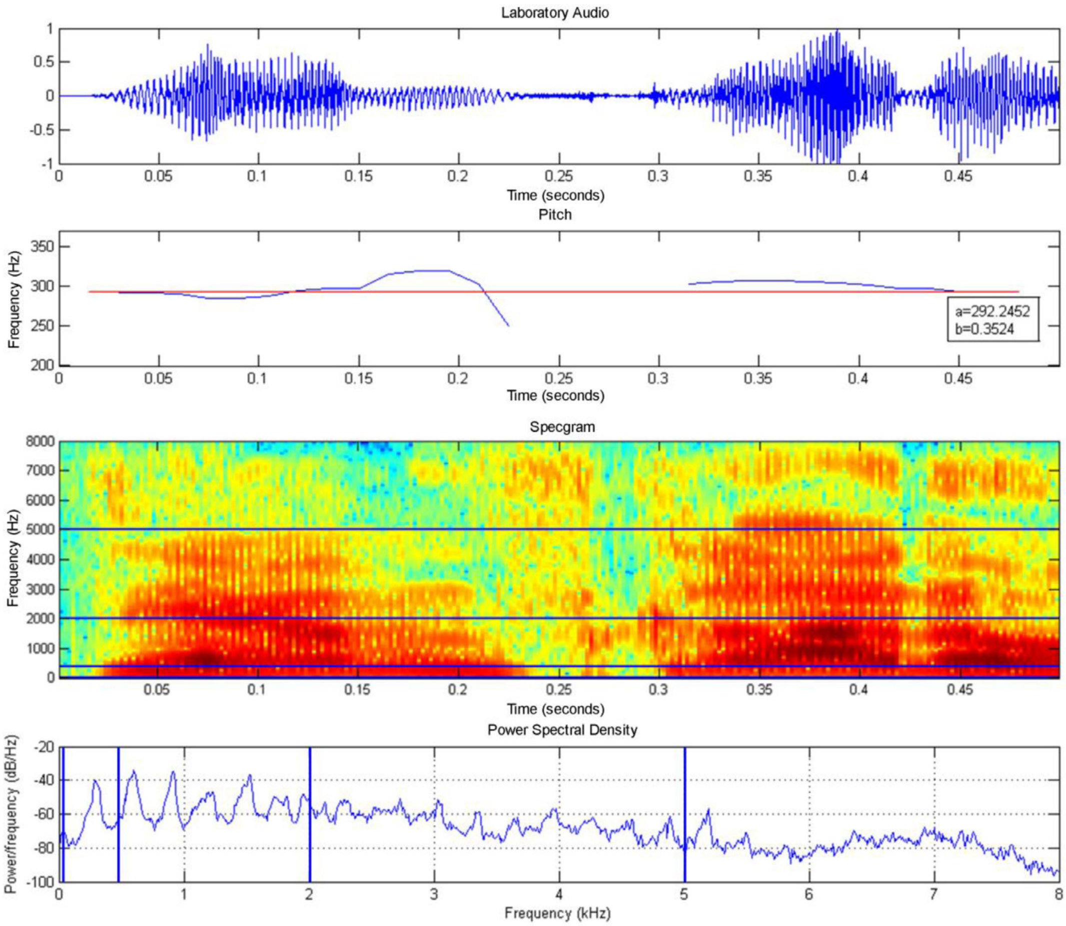

Fig. 2. Prosodic and paralinguistic features.

they are quickly and easily calculated, secondly, their robustness in emotion recognition has been proven, and finally, they are independent of linguistic segmentation.

2.3.1. Prosodic features

Fundamental frequency is the main prosodic indicator; to be more specific, the intonation of speech (pitch contour). In general, the pitch contour of a neutral speech locution starts with the maximum value of pitch and from that point it starts falling until the final value, so that the contour has a slightly descending global slope. In case of emotional locutions this global slope changes. This circumstance is more significant with different activation levels.

In this paper we use two prosodic features: the two linear regression coefficients (a and b) obtained from modeling the pitch contour *p(n)* (Kwon et al., 2003; Petrushin, 1999) from $\{w(n)\}$ following Eq. (1):

$$MIN(a,b) = \sum_{i=1}^{n} (p_i(n) - a - bx_i(n))^2 \quad (1)$$

where the coefficients a and b are computed using the method of least squares (Fig 2). The coefficient a represent the original pitch and the coefficient b is related with the decline or trend of the pitch. We use the pitch estimation algorithm called YIN (De Cheveigné & Kawahara, 2002) in our implementation.

2.3.2. Paralinguistic features

Short-term energy is a prosodic indicator, which is used to convey the lexical tension in speech. The accumulation of voiced energy in different frequency bands, which vary depending on the speech production model, can also be used as a paralinguistic indicator of the emotional state. The high-frequency energy increases in emotional speech compared to the neutral version.

In this paper we use four paralinguistic features which are four voiced spectral energy balances (E_{B_0},E_{B_1}, E_{B_2}, and E_{B_3}) calculated from each voiced frame $\{w(n)\}$. They are quantified using 4 percentages of energy concentration in 4 frequency bands B_i (where $i \in [0, 3]$). For a sampling frequency greater than 16 kHz, the frequency bands are divided into the following ranges: B_0=[0 Hz, 400 Hz], B_1=[400 Hz, 2 kHz], B_2=[2 kHz, 5 kHz], and B_3=[5 kHz, 8 kHz] (Fig. 2). These bands have been estimated in our previous research related to the phonatory system (Alonso, De Leon, Alonso, & Ferrer, 2001). The percentage of energy in each frequency band E_{B_i}, Eq. (2), is obtained using the following expression:

$$E_{B_i} = \frac{\sum\limits_{f=B_i} |X(f)|^2}{\sum\limits_{f=0}^{8KHz} |X(f)|^2} \quad 0 \le i \le 3 \quad (2)$$

Where $|X(f)|^2$is a periodogram of the temporal voiced frame$\{w(n)\}$.

2.4. Classifiers

Support vector machines (SVM) (Burges, 1998) have been used in this study for implementing the measure *Emotional Temperature*. We have employed a freely available implementation named LIBSVM (Chang & Lin, 2011) with radial basis kernel function that has been

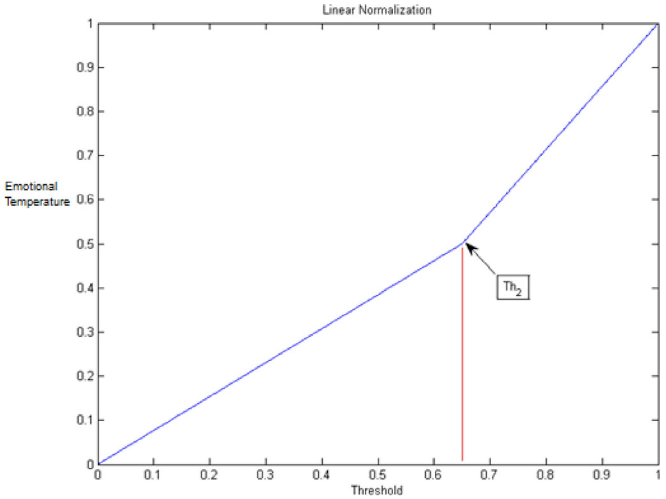

Fig. 3. Linear normalization of percentage of voiced frames labeled as emotional.

used successfully in other research (Chavhan et al., 2015; Fan et al., 2014; Harimi et al., 2014).

2.5. Emotional temperature

A simple hierarchical approach in two stages is carried out. Previously the speech signal has been temporarily segmented into emotional segments and the proposed feature set has been extracted from each one of them.

Firstly, each emotional segment is classified from LIBSVM in two classes: high activation and low activation. A decision threshold (Th_1) is calculated from the Equal Error Rate (EER) obtained from the training data.

Secondly, the speech signal is classified as high activation when the percentage of emotional segments classified as high activation in the first stage in relation with the whole emotional segments analyzed from the speech is located above a second threshold (Th_2). The second threshold (Th_2) is calculated from the EER obtained from the validation data to estimate the minimum percentage of high activation segments necessary to classify the speech signal as high activation.

The resulting scale from this framework is linear normalized, in relation to the percentage of emotional segment classified as high activation from the speech signal, in a way that takes the value of 0 in the 0%, 50 in the Th_2 and 100 for 100% (Fig 3). This normalized scale is called *Emotional Temperature (ET)* and allows a progressive assessment of the activation level detected in continuous speech, where a speech signal is classified as high activation when ET $\geq$50 and as low activation in the contrary case.

2.6. Evaluation

Three experiments have been made with text-independent and speaker-independent. Each experiment evaluates the databases EMO-DB, LDC-DB and Polish database separately and distinguishing by genders (male and female). The first experiment analyzes the proposed feature set (prosodic and linguistic features), its ability to recognize high activation and the detection threshold Th_1 that is set by EER.

The second experiment analyzes the average EER and detection threshold Th_2 for percentages of emotional segments classified as high activation.

Finally, the third experiment shows the success rates and several performance parameters for the three emotional databases using the *Emotional Temperature* method and compares the results obtained with others researches. In addition, the activation level for five emotions in the *Emotional Temperature* scale is analyzed.

To increase the reliability of experiments, each study is repeated 10 times ($k = 1, \ldots, 10$), the samples for each class are uniformly distributed and in each iteration the database is randomly divided into 40% samples for training (training set), 20% samples for optimization of the SVM classifiers (optimization set), 20% samples for validate the threshold Th_2 (validation set) and 20% samples for test (testing set).

The data in the training set are z-score normalized. The training set is normalized by subtracting the training set mean and dividing by the training set standard deviation for each feature. The testing set, validation set and optimization set are normalized according to the normalization values used for the training set.

3. Results & discussion

3.1. Analysis of the proposed feature set

In this subsection, the proposed feature set is analyzed. The analysis is carried out for five of the emotions more recorded and analyzed in the state-of-art (Ververidis & Kotropoulos, 2003): anger, boredom, happy, neutral state and sadness. The emotions anger and happy represent high activation, and the emotions neutral state, boredom and sadness represent low activation. Figs. 4–5 show the distribution, using box plots, of the six features (prosodic and paralinguistic features) for the Berlin emotional speech database, for the LDC emotional speech database and for the Polish emotional speech database, distinguishing by genders.

According to the Fig 4 the median of the values of the prosodic feature *a* is higher in anger and happy emotional speech than the rest of emotional speech, showing its relationship with the high activation where the pitch is higher. This tendency is the same in the three databases and for the two genders. In the case of the prosodic feature *b*, which is related with the decline or trend of the pitch, does not have a clear differences between the five emotions in our experiment. The median of the distributions are overlapped for this feature in the case of the LDC database and the Berlin database.

In the Fig 5 the spectral energy balances are analyzed. The median of the values of the voice spectral energy balances (paralinguistic feature) in medium and high frequencies (400 Hz–8 kHz) is higher in anger and happy emotional speech than the rest of emotional speech. Likewise, the median of the values of the voice spectral energy balance in low frequencies (60–400 Hz) is lower in anger and happy emotional speech than the rest of emotional speech. In both cases, the difference between the values of the voice spectral energy balances in the group of high activation –anger and happy emotional speeches- and in the group of low activation –boredom, neutral and sadness emotional speeches– are clear in the three databases and for the two genders. This is an indicator of the voice spectral energy balances are a good feature to discriminate emotional speech.

According to the data distributions, the six features are discriminative between high activation. The data distributions are very similar in the three databases and for the two genders. This suggests that the discriminatory ability of the proposed feature set is prior language-independent, although in this first approach we have not done the experiment showing the cultural independence of the feature set.

For each emotional segment, the six features were evaluated with LIBSVM for the three emotional speech databases (Berlin, LDC and Polish databases) and by genders using two emotions limits: anger emotional speech (high activation) and sadness emotional speech (low activation).

The EER and the detection threshold Th_1 for optimum recognition of emotional segments of high activation are showed in the Fig. 6 and the Table 1. The EER and the detection threshold are

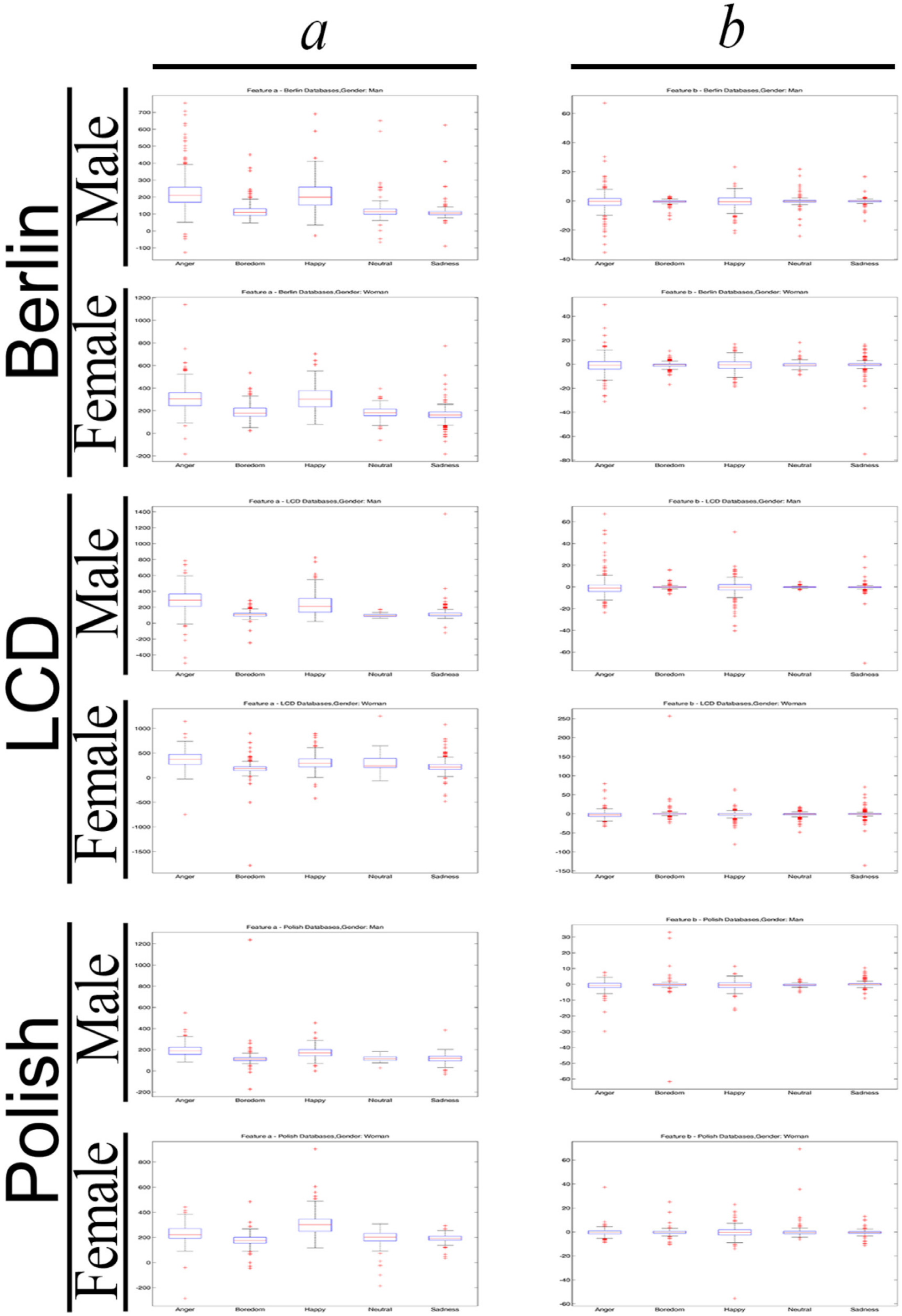

Fig. 4. Quartile study of prosodic feature (*a* and *b*) for five emotions.

obtained averaging the results of each interaction. According to the results, the measures shows a good discriminatory ability between two emotional states (high activation and low activation), with a low EER in the three databases and the two genders, still more significant to the male gender.

3.2. Analysis of percentages emotional segments in the speech signal

In this subsection, the percentages of emotional segments classified as high activation in a speech are analyzed. The second detection threshold Th_2 is fixed from the validation data to estimate the

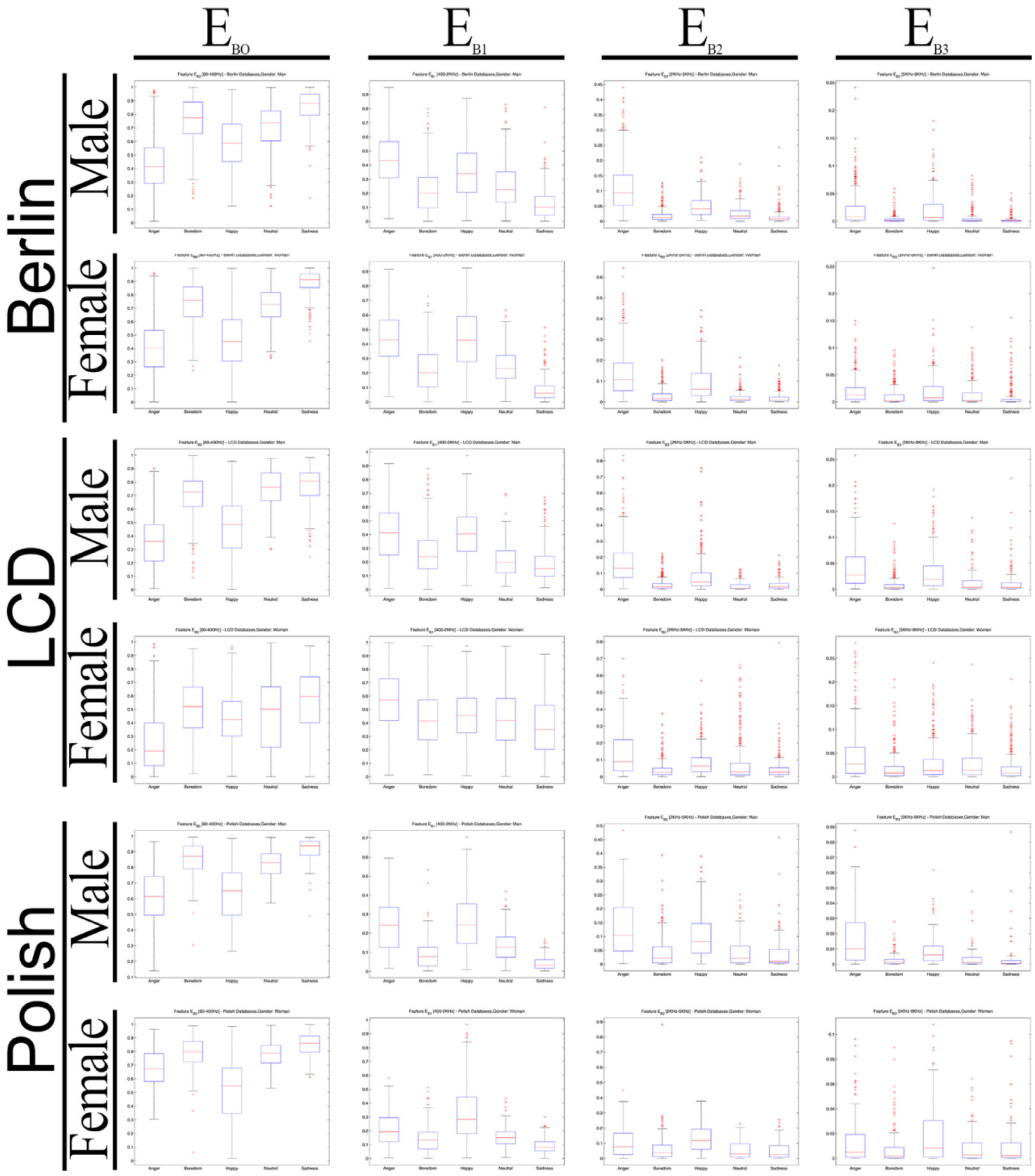

Fig. 5. Quartile study of paralinguistic feature (E_{BO}, E_{B1}, E_{B2}, E_{B3}) for five emotions.

percentage of emotionally salient segment which are classified as high activation, in the first stage, in relation to the total emotional segments of the speech signal. Th_2 is determined and the analysis is done in the three databases and by genders for two emotional states: high activation, represented by anger emotion, and low activation, represented by sadness emotion.

The Fig. 7 shows the Receiver Operating Characteristic (ROC) curve. In this ROC curve the true high activation segments are plotted in function of the false high activation segments. The Area Under the ROC curve (AUC) is a measure of how well can distinguish between high activation and low activation. In the three database, for the male gender the AUC has a value close to 1, being indicator that provides a discrimination value nearly perfect, while that for the female gender, the AUC has a value of close to 0.75 , indicated is a good discrimination test.

The Table 2 shows the average EER and the optimal detection threshold Th_2 for emotional segments recognition of high activation. The EER and this threshold are obtained averaging the results of each interaction. According to the results, the measure shows a good discriminatory ability between high activation and low activation, with a low EER in the three databases for the male gender and a tolerable EER for the female gender.

3.3. *Results of the databases evaluation with* Emotional Temperature

The *Emotional Temperature* was evaluated for the three emotional speech databases (Berlin, LDC and Polish databases) and the results are compared with other similar researches. The first analysis is done for five emotions: anger, boredom, happy, neutral state and sadness. The mean value of *Emotional Temperature* for the different emotions and the different databases are shown in Table 3 with the standard deviation (σ). The estimation of value of *Emotional Temperature* rates is obtained averaging the results of each iteration. According to the results, the *Emotional Temperature* value is clearly higher in anger and

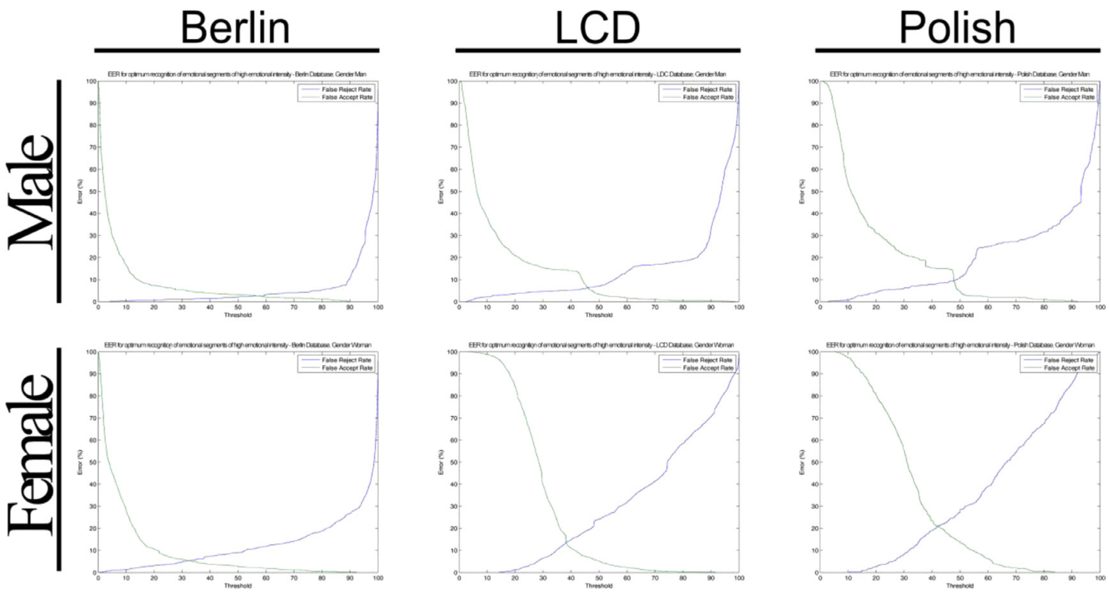

Fig. 6. EER for optimum recognition of emotional segments of high activation.

Table 1
Average EER and detection threshold Th_1 for optimum recognition of emotional segments of high activation.

Gender/Database	Male		Female	
	EER (%)	Th_1 (%)	EER (%)	Th_1 (%)
Berlin	2.58	56.46	5.34	32.44
LCD	6.27	46.18	13.31	38.16
Polish	9.35	47.81	20.86	42.13

Table 2
Average EER and detection threshold Th_2 for the percentages of segments classified as high activation.

Gender/Database	Male		Female	
	EER (%)	Th_2 (%)	EER (%)	Th_2 (%)
Berlin	0.000	65.005	0.000	51.285
LCD	3.846	70.835	19.333	58.575
Polish	2.500	52.085	27.500	49.495

happy emotional speeches than boredom, neutral and sadness emotional speeches. Therefore, the *Emotional Temperature* shows good discriminatory ability of activation level in the three databases.

Furthermore, the performance of *Emotional Temperature* in the automatic speech emotion recognition between high and low activation has been evaluated. The Table 4 shows the system performance analysis. Seven parameters are evaluated: accuracy (measures the reliability of the system for a high activation), sensitivity (ratio of right answers in genuine comparisons respect the total of comparisons. It measures the ability of the system to detect the high activation tested), specificity (ratio of right answers in non-genuine comparisons respect the total of comparisons. As higher the value is, more difficult is to replace the high activation by another), Positive Predictive Value (PPV) (a measure of the probability of a true positive result is a true positive), Negative Predictive Value (NPV) (it is a measure of the probability that a negative result is really a true negative), False Positive Rate (FPR) (false acceptance ratio to the total of non-genuine comparisons. It represents the validity of the system) and False Discovery Rate (FDR) (false acceptance ratio to the total of genuine comparisons. FDR procedures are designed to control the expected proportion of incorrectly accepted false positives).

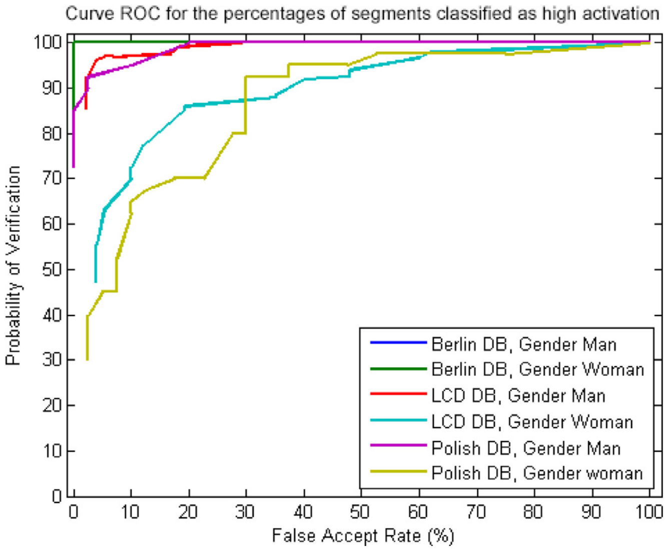

Fig. 7. Curve ROC for the percentages of segments classified as high activation.

Table 3
Emotional Temperature by emotions.

Database	Emotion input	Male	Female
Berlin	Anger	97.68 ($\sigma = 6.26$)	95.84 ($\sigma = 8.36$)
	Boredom	22.62 ($\sigma = 21.08$)	37.9 ($\sigma = 21.37$)
	Happy	92.07 ($\sigma = 12.63$)	90.69 ($\sigma = 13.89$)
	Neutral	30.97 ($\sigma = 28.01$)	46.04 ($\sigma = 25.55$)
	Sadness	4.98 ($\sigma = 7.28$)	4.41 ($\sigma = 7.01$)
LCD	Anger	92.89 ($\sigma = 19.20$)	86.57 ($\sigma = 19.68$)
	Boredom	7.95 ($\sigma = 17.3$)	24.84 ($\sigma = 22.06$)
	Happy	68.95 ($\sigma = 35.43$)	54.65 ($\sigma = 30.3$)
	Neutral	4.2 ($\sigma = 14.81$)	42.74 ($\sigma = 35.65$)
	Sadness	8.3 ($\sigma = 19.45$)	17 ($\sigma = 21.76$)
Polish	Anger	93.58 ($\sigma = 12.3$)	81.01 ($\sigma = 22.79$)
	Boredom	23.76 ($\sigma = 22.71$)	51.75 ($\sigma = 24.67$)
	Happy	87.34 ($\sigma = 15.88$)	95.19 ($\sigma = 6.16$)
	Neutral	33.73 ($\sigma = 31.15$)	64.63 ($\sigma = 18.98$)
	Sadness	5.85 ($\sigma = 9.24$)	37.65($\sigma = 24.04$)

Table 4
Study of predictive capacity.

		Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	FPR (%)	FDR (%)
Berlin	Male	94.9	100	89.8	90.74	100	10.2	9.26
	Female	85.77	100	71.55	77.85	100	28.46	22.15
LCD	Male	88.32	82.58	94.06	93.29	84.37	5.94	6.71
	Female	73.89	67.2	80.57	77.573	71.07	19.43	22.43
Polish	Male	90	100	80	83.34	100	20	16.67
	Female	67.5	95	40	61.29	88.89	60	38.71

Table 5
Comparison of activation recognition results in researches emotions: anger (A), boredom (B), disgust (D), fear (F), happy (H), neutral (N), sadness (S). Accuracy: Male/Female (M/F), Gender-Independent (GI).

Research	Database	Emotions	Feature set	Classification method	Accuracy (%)
(Lika et al., 2014)	EMO-DB	A, H, N, S	41 features	ANFC	98.52 (GI)
	RML-DB				92.30 (GI)
(Iriya & Arjona Ramirez, 2014)	EMO-DB	A, B, H, N, S, F, D	8 features	GMM	93.77–96.43 (M/F)
Emotional Temperature	EMO-DB	A, B, H, N, S	6 features	SVM	94.9–85.77 (M/F)
	LDC-DB				88.32–73.89 (M/F)
	Polish-DB				90–67.5 (M/F)

The results show that the *Emotional Temperature* has good discriminatory ability between high and low activation, with an average accuracy that range from 94.9% for male in the Berlin database to 67.5% for female in the Polish database. In all database the proposed approach has a better performance in the male gender with a total average accuracy of 91.07%.

Comparison of results in the state-of-art is a difficult task, since most often the databases, the conditions of dependence/independence, the emotions and the number of emotions that represent the activation space, the experiments and the representation of the results are different. However, in the Table 5 we have tried to make a comparison with other similar researches in the field of activation recognition.

The comparative researches used in the common the EMO-DB database, and do not differ in excess in the emotions and the number of emotions employed. In these conditions, our approach shows similar accuracy rate using a feature set with less features and lower complexity compared to the rest of research which used MFCC and LFPC. As a result, the proposed approach provides a good method to activation level recognition in this field.

4. Conclusions

This paper focuses on the complex task of quantified activation from emotional speech. The approach has been developed with the goal of being able to attend in the future development of real-time applications in automatic speech emotion recognition and continuous tracking systems of user emotion capable of detecting changes in the speaker's emotional state.

In this paper, activation recognition is implemented to monitor user emotional states by trying to distinguish between two categories: high activation, represented by anger and happy emotions, and low activation, represented by neutral, boredom and sadness emotions. A temporal segmentation and emotion recognition of each segment is used compared to the classical approach used in other researches that assume that each analyzed utterance contains only one emotional state. The temporal segmentation allows the advantage of assisting the detection emotional turns to carry out a continuous evaluation from the speech signal. The proposed feature set include six features (two prosodic features and four paralinguistic features), which are robust, low complexity and high speed computing, ideals to develop real-time applications. Our proposed method was researched on three databases and the results showed that it could be effectively applied in acted corpus to discriminate between high and low activation through a hierarchical segmentation, feature extraction and classification system that provide a scale for measuring of high/low activation, called *Emotional Temperature*. In comparison with other similar researches, *Emotional Temperature* has been shown to have an accuracy rate as good as the rest of researches with the advantage of using a smaller and simpler feature set.

In spite of this, first approach to the quantification of activation has been studied with acted speech in terms of language and gender dependence showing a lower accuracy rate in female speech, as well as a higher variation in accuracy rate between languages for this gender.

In future work, the approach performance should be improved for the female speech, an analysis of more profound cultural independence must be carried out in search of ensure a universal emotion system; real-life databases will be attempted. We will also consider study the *Emotional Temperature* applicability in affective speech interface for users with mental disorders or neurodegenerative disease, such as Parkinson's disease or Alzheimer's disease, which are characterized by monotony of pitch and loudness, reduced stress, variable rate, imprecise consonants, and a breathy and harsh voice, all of which affect a patient's ability to produce an emotional tone of voice.

With all this, the proposed approach provides a comparable performance with lower complexity than other approaches for real-time applications, thus making it an appealing alternative, may assist in the future development of automatic speech emotion recognition systems with continuous tracking, and was its potential applications the detection of the emotional state in telephone call-center conversations; diagnostic and therapeutic tools for psychology, psychiatry and neurology where is needed detect variations in the patient's emotional state or their inability to manifest it; adaptive learning environments to the student's emotional state and in general, any human computer interaction system, which is required to recognize the human emotion as it is, continuum.

References

Alepis, E., & Virvou, M. (2011). Automatic generation of emotions in tutoring agents for affective e-learning in medical education. *Expert Systems with Applications, 38*(8), 9840–9847.

Ali, H., Hariharan, M., Yaacob, S., & Adom, A. H. (2015). Facial emotion recognition using empirical mode decomposition. *Expert Systems with Applications, 42*(3), 1261–1277.

Alonso, J. B., De Leon, J., Alonso, I., & Ferrer, M. A. (2001). Automatic detection of pathologies in the voice by HOS based parameters. *EURASIP Journal on Applied Signal Processing, 4*, 275–284.

Altun, H., & Polat, G. (2009). Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. *Expert Systems with Applications, 36*(4), 8197–8203.

Ammar, M. B., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The affective tutoring system. *Expert Systems with Applications, 37*(4), 3013–3023.

Amol, T. K., & Guddeti, R. M. R. (2014). Multiclass SVM-based language-independent emotion recognition using selective speech features. In *Proceedings of ICACCI International Conference on Advances in Computing, Communications and Informatics* (pp. 1069–1073).

Balti, H., & Elmaghraby, A. S. (2014). Emotion analysis from speech using temporal contextual trajectories. In *Proceedings of IEEE Symposium on Computers and Communication (ISCC), 2014* (pp. 1–7).

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. In *Proceedings of IEEE Transactions on Acoustics, Speech and Signal Processing: 27* (pp. 113–120).

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech, 5*, 1517–1520.

Cabrera, J., Alonso, J. B., Travieso, C. M., Ferrer, M. A., Hernriquez, P., Dutta, M. K., et al. (2015). Emotional states discrimination in voice in secure environments. In *Proceedings of the 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015* (pp. 843–847). doi:10.1109/SPIN.2015.7095414.

Chang, C., & Lin, C. (2011). LIBSVM: a library for support vector machines. In *Proceedings of ACM Transactions on Intelligent Systems and Technology (TIST): 2* (p. 27).

Chavhan, Y., Yelure, B., & Tayade, K. (2015). Speech emotion recognition using RBF kernel of LIBSVM. In *Proceedings of the 2nd International Conference on Electronics and Communication Systems (ICECS), 2015* (pp. 1132–1135).

Chenchah, F., & Lachiri, Z. (2014). Speech emotion recognition in acted and spontaneous context. *Procedia Computer Science, 39*, 139–145.

Coutinho, E., Deng, J., & Schuller, B. (2014). Transfer learning emotion manifestation across music and speech. In *Proceedings of International Joint Conference on Neural Networks (IJCNN), 2014* (pp. 3592–3598).

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40*(1), 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE, 18*(1), 32–80.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71*, 10–49.

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917–1930.

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*(16), 6351–6358.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Fan, Y., Xu, M., Wu, Z., & Cai, L. (2014). Automatic emotion variation detection in continuous speech. In *Proceedings of Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)* (pp. 1–5).

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America, 128*(3), 1322–1336.

Harimi, A., Shahzadi, A., & Ahmadyfard, A. (2014). Recognition of emotion using nonlinear dynamics of speech. In *Proceedings of the 7th International Symposium on Telecommunications (IST), 2014* (pp. 446–451).

Iriya, R., & Arjona Ramirez, M. (2014). Gaussian mixture models with class-dependent features for speech emotion recognition. In *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP), 2014* (pp. 480–483).

James, W. (1884). II.—What is an emotion? *Mind*, (34), 188–205.

Julia, F. N., & Iftekharuddin, K. M. (2005). Detection of emotional expressions in speech. In *Proceedings of the IEEE, SoutheastCon, 2006* (pp. 307–312).

Kamaruddin, N., Wahab, A., & Quek, C. (2012). Cultural dependency analysis for understanding speech emotion. *Expert Systems with Applications, 39*(5), 5115–5133.

Kanagaraj, S. A., Shahina, A., Devosh, M., & Kamalakannan, N. (2014). EmoMeter: measuring mixed emotions using weighted combinational model. In *Proceedings of International Conference on Recent Trends in Information Technology (ICRTIT), 2014* (pp. 1–6).

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., et al. (2014). Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014* (pp. 4858–4862).

Kitayama, S. E., & Markus, H. R. E. (1994). Emotion and culture: empirical studies of mutual influence. *American Psychological Association*, 1–385.

Kostoulas, T., Mporas, I., Kocsis, O., Ganchev, T., Katsaounos, N., Santamaria, J. J., et al. (2012). Affective speech interface in serious games for supporting therapy of mental disorders. *Expert Systems with Applications, 39*(12), 11072–11079.

Kwon, O., Chan, K., Hao, J., & Lee, T. (2003). Emotion recognition by speech signals. *Interspeech*.

Laukka, P. (2004). Vocal expression of emotion: discrete-emotions and dimensional accounts. *Acta Universitatis Uppsalensis, Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, 141*, 1–80.

Le, B. V., & Lee, S. (2014). Adaptive hierarchical emotion recognition from speech signal for human-robot communication. In *Proceedings of 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014* (pp. 807–810).

Lee, C. M., & Narayanan, S. (2003). Emotion recognition using a data-driven fuzzy inference system. *INTERSPEECH*.

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications, 41*(4), 1742–1749.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). *Emotional Prosody Speech And Transcripts*. P hiladelphia: Linguistic Data Consortium.

Lika, R. A., Seldon, H. L., & Kiong, L. C. (2014). Feature analysis of speech emotion data on arousal-valence dimension using adaptive neuro-fuzzy classifier. In *Proceedings of International Conference on Industrial Automation, Information and Communications Technology (IAICT), 2014* (pp. 104–110).

Lopez-de-Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., et al. (2013). On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation, 7*(1), 44–55.

López-Ludeña, V., Barra-Chicote, R., Lutfi, S., Montero, J. M., & San-Segundo, R. (2013). LSESpeak: a spoken language generator for deaf people. *Expert Systems with Applications, 40*(4), 1283–1295.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., et al. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems, 63*, 68–81.

Mustafa, M. B., & Ainon, R. N. (2013). Emotional speech acoustic model for malay: iterative versus isolated unit training. *The Journal of the Acoustical Society of America, 134*(4), 3057–3066.

Nass, C., Jonsson, I., Harris, H., Reaves, B., Endo, J., Brave, S., et al. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 1973–1976.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: a systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Neiberg, D., & Elenius, K. (2008). Automatic recognition of anger in spontaneous speech. *INTERSPEECH*, 2755–2758.

Ooi, C. S., Seng, K. P., Ang, L., & Chew, L. W. (2014). A new approach of audio emotion recognition. *Expert Systems with Applications, 41*(13), 5858–5869.

Oudeyer, P. (2002). Novel useful features and algorithms for the recognition of emotions in human speech. In *Proceedings of Speech Prosody 2002, International Conference*.

Pao, T., Chien, C. S., Chen, Y., Yeh, J., Cheng, Y., & Liao, W. (2007). Combination of multiple classifiers for improving emotion recognition in mandarin speech. In *Proceedings of 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2007: 1* (pp. 35–38).

Pao, T., Chien, C. S., Yeh, J., Chen, Y., & Cheng, Y. (2007). Continuous tracking of user emotion in mandarin emotional speech. In *Proceedings of the 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2007: 1* (pp. 47–52).

Petrushin, V. (1999). Emotion in speech: recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering* (pp. 7–10).

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application., 222–225.

Picard, R. (1999). Affective computing for HCI. *Human-Computer Interaction: Ergonomics and User Interfaces, 1*, 829–833.

Pohjalainen, J., & Alku, P. (2014). Multi-scale modulation filtering in automatic detection of emotions in telephone speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014* (pp. 980–984).

Poon-Feng, K., Huang, D., Dong, M., & Li, H. (2014). Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines. In *Proceedings of 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2014* (pp. 584–588).

Rabiei, M., & Gasparetto, A. (2014). A system for feature classification of emotions based on speech analysis; applications to human-robot interaction. In *Proceedings of International Conference on Robotics and Mechatronics (ICRoM), 2014 Second RSI/ISM* (pp. 795–800).

Razak, A. A., Komiya, R., Izani, M., & Abidin, Z. (2005). Comparison between fuzzy and NN method for speech emotion recognition. In *Proceedings of 3rd International Conference on Information Technology and Applications, ICITA 2005: 1* (pp. 297–302).

Roach, P. (2000). Techniques for the phonetic description of emotional speech. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 53–59).

Roy, D., & Pentland, A. (1996). Automatic spoken affect classification and analysis. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, 1996.* (pp. 363–367).

Russell, J. A. (1997). How shall an emotion be called?

Sankar, R. (1988). Pitch extraction algorithm for voice recognition applications. In *Proceedings of the 20th Southeastern Symposium on System Theory, 1988.* (pp. 384–387).

Scherer, K. R. (1981). Speech and emotional states. *Speech Evaluation in Psychiatry*, 189–220.

Scherer, K. (1987). Toward a dynamic theory of emotion: the component process model of affective states. *Geneva Studies in Emotion and Communication, 1*, 1–98.

Schirmer, A., Striano, T., & Friederici, A. D. (2005). Sex differences in the preattentive processing of vocal emotional expressions. *Neuroreport, 16*(6), 635–639.

Staroniewicz, P., & Majewski, W. (2009). Polish emotional speech database–recording and preliminary validation. *Cross-modal analysis of speech, gestures, gaze and facial expressions* (pp. 42–49). Springer.

Tawari, A., & Trivedi, M. (2010). Speech emotion analysis in noisy real-world environment. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR), 2010* (pp. 4605–4608).

Ververidis, D., & Kotropoulos, C. (2003). A state of the art review on emotional speech databases. In *Proceedings of the 1st Richmedia Conference* (pp. 109–119).

Vidrascu, L., & Devillers, L. (2007). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In *Proceedings of International Workshop on Paralinguistic Speech between Models and Data, ParaLing*.

Vogt, T., & André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of IEEE International Conference on Multimedia and Expo, ICME 2005.* (pp. 474–477).

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., et al. (2008). Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. *INTERSPEECH*, 597–600 2008.

Wöllmer, M., Schuller, B., Eyben, F., & Rigoll, G. (2010). Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *Selected Topics in Signal Processing, IEEE Journal Of, 4*(5), 867–881.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2007). Automatic hierarchical classification of emotional speech. In *Proceedings of the 9th IEEE International Symposium on Multimedia Workshops, ISMW'07.* (pp. 291–296).

Yacoub, S. M., Simske, S. J., Lin, X., & Burns, J. (2003). Recognition of emotions in interactive voice response systems. *INTERSPEECH*.

Zhang, Y., Zhang, L., & Hossain, M. A. (2015). Adaptive 3D facial action intensity estimation and emotion recognition. *Expert Systems with Applications, 42*(3), 1446–1464.

Zhao, S., Rudzicz, F., Carvalho, L. G., Márquez-Chin, C., & Livingstone, S. (2014). Automatic detection of expressed emotion in parkinson's disease. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014* (pp. 4813–4817).