Int J Speech Technol
DOI 10.1007/s10772-012-9127-7

# Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema

**Margarita Kotti · Fabio Paternò**

Received: 12 September 2011 / Accepted: 12 January 2012
© Springer Science+Business Media, LLC 2012

**Abstract** In this paper, a psychologically-inspired binary cascade classification schema is proposed for speech emotion recognition. Performance is enhanced because commonly confused pairs of emotions are distinguishable from one another. Extracted features are related to statistics of pitch, formants, and energy contours, as well as spectrum, cepstrum, perceptual and temporal features, autocorrelation, MPEG-7 descriptors, Fujisaki's model parameters, voice quality, jitter, and shimmer. Selected features are fed as input to $K$ nearest neighborhood classifier and to support vector machines. Two kernels are tested for the latter: linear and Gaussian radial basis function. The recently proposed speaker-independent experimental protocol is tested on the Berlin emotional speech database for each gender separately. The best emotion recognition accuracy, achieved by support vector machines with linear kernel, equals 87.7%, outperforming state-of-the-art approaches. Statistical analysis is first carried out with respect to the classifiers' error rates and then to evaluate the information expressed by the classifiers' confusion matrices.

**Keywords** Emotion recognition · Large-scale feature extraction · Binary classification schema · Speaker-independent protocol · Classifier comparison

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

M. Kotti (✉) · F. Paternò
ISTI-CNR, Via G. Moruzzi, 1, 56124 Pisa, Italy
e-mail: margarita.kotti@isti.cnr.it

F. Paternò
e-mail: fabio.paterno@isti.cnr.it

Published online: 31 January 2012

## 1 Introduction

Human behavior is a natural reference for artificial systems. Psychology and neurology suggest that emotions are important in decision making, problem solving, cognition, and intelligence. The vision of future computing is human-centered (Pantic et al. 2006), and should take affect into account (Picard 1997). Future computing will be characterized by its ease of use and it should adapt automatically to the users' behavioural patterns. Within this context, emotion recognition is considered to be a fundamental aspect for human-computer interaction (HCI). Emotion recognition could provide users with improved services by being adaptive to their emotions. For example, an angry user should be appraised, a confused user might be offered an alternative explanation by the computer, whereas the system should also be capable to share the happiness to corroborate the partnership between user and system.

Vocal expression is a primary carrier of affective signals in human communication (Nass et al. 2005). Today there is a need to know not only what the user says, but also how he/she says it (Lee and Narayanan 2005). The aim is to improve naturalness and efficiency of spoken human-machine interfaces (Cowie et al. 2001), through the exploitation of paralinguistic properties (Yang and Lugger 2010).

However, emotion recognition is a challenging task, even for humans (Chandaka et al. 2009; Lee and Narayanan 2005), as is verified by the related literature (Bosma and André 2004; Fersini et al. 2009; Iliou and Anagnostopoulos 2009). This can be attributed to a multitude of reasons. To begin with, emotions are difficult to define from a psychological point of view (Lee and Narayanan 2005). In fact, emotions are ill-defined, possibly indeterminate, and they exhibit fuzzy boundaries that cannot be directly measured (Calvo and D'Mello 2011). There are ongoing debates about how many emotion categories exist (Lee and

Narayanan 2005). In addition, there may be more than one perceived emotion in the same utterance. There is a lack of definite description and agreement upon one set of basic universal emotions (Lee and Narayanan 2005). For example, remorse is a combination of sadness and disgust (Fersini et al. 2009). In addition, different emotions can share similar properties. In (Chandaka et al. 2009), it is stated that emotional signals are chaotic in nature whereas it is a fact that researchers have not yet found a sufficient feature set to describe the emotional states efficiently (Fersini et al. 2009). It is also proven that each speaker expresses his/her emotions in a different manner (Fersini et al. 2009) and that the two genders convey their emotions in profoundly different ways (Fersini et al. 2009; Nass et al. 2005; Pittermann et al. 2010; Ververidis and Kotropoulos 2005; Zeng et al. 2007). Furthermore, the same linguistic content may bear different emotional state.

Emotion recognition can find several applications. For the case of intelligent assistance, emotion recognition can increase usability, user-friendliness, and cooperativeness by adapting dynamically to the emotional state of the user (Minker et al. 2007). This way the interaction is more natural, simplified and accelerated. Another potential application refers to call centres management. When a customer experiences negative emotions the system can either adjust to the customer needs or pass the control to a human agent. It is proven that the emotion of an automobile voice can affect the driver's performance, leading to less accidents (Nass et al. 2005). Also, fatigue can be detected in the driver's voice and the driver can be alerted (Chandaka et al. 2009). Other examples of emotion aware systems include support for people with disabilities, such as educational software (Bosma and André 2004) for people with autism (Konstantinidis et al. 2009) or serious visual impair. Alternative applications deal with emotion detection in games and human-robot interaction. Surveillance and detection of potentially hazardous events are also possible applications of emotion recognition (Ntalampiras et al. 2009), whereas voice mail refinement could be benefited, as well (Inanoglu and Caneel 2005). Other possible applications include commercial products, life-support systems, virtual guides, customer service, lie detectors, conference room research, emotional speech synthesis, art, entertainment etc.

Additionally, emotion recognition has found alternative uses in health care systems. For example, anger and impatience is detected in a database of speech recordings derived during surgery operations (Schuller et al. 2008). Thus, in the case of an angry or impatient surgeon a security action can be carried out. Psychiatry may use emotion recognition to attend patients with psychological problems, such as depression, deception or anxiety (Ekman et al. 2005; Hirschberg et al. 2005). Additional health science areas include behavioral science and neuroscience (Zeng et al. 2007) by the development of tools, which are capable to improve the reliability of measurements and accelerate the tedious and demanding task of manually processing data on human affective behavior.

Our system exploits exclusively the audio channel. It comprises 3 modules: feature extraction, feature selection, and classification. The current paper begins with a review of the related literature. Then, the *first contribution* is presented, which is a *psychologically-inspired binary cascade classification schema*. The schema adopts the mental dimensional descriptors of valence, activation, and stance (Ekman and Davidson 1994). It decomposes the multi-class emotion recognition problem into several binary classification problems. The reason is that a feature that presents a good discrimination ability for a set of emotions may not present the same attribute for an alternative set of emotions. Other advantages include better separation of commonly confused emotional states, easy adaptation to databases with diverse emotions, and the possibility to choose the stop level of the schema. A *second contribution* lies in the extraction of a large pool of 2327 features from the Berlin emotional speech database (EMODB). Several features, namely MPEG-7 descriptors, Teager energy operator on autocorrelation, total loudness, and specific loudness sensation coefficients (SLSC), which all told amount to *602 novel features*, are proposed in this study for the first time to the best of the authors knowledge within the context of emotion recognition. The normalized extracted features are fed as input to a forward feature selection algorithm for each gender separately. Next, speaker-independent experiments are carried out for *each gender separately*, which constitutes the *final contribution* of the paper. Speaker-independency is a new trend in the research community and consequently a limited number of speaker-independent contributions is available. Speaker-independent systems are more robust and stable and they demonstrate a better generalization ability than the speaker-dependent ones. Furthermore, speaker-independent systems are ideal when the number of speakers is limited. As classifiers Support vector machines (SVMs) as well as $K$ nearest neighborhood ($K$NNs) are applied and efficiency is presented by means of confusion matrices as well as emotion recognition accuracy. Two different kernel functions are tested for SVMs, namely linear and Gaussian radial basis function. Performance analysis of $K$NN as well as SVM with Gaussian radial basis function kernel under diverse parametrization is performed. An amendment of the paper is related to the statistical analysis of the experimental results. In particular, on the one hand the classifiers' error rates are compared, while on the other hand the confusion matrices are ranked. The final advantage is the comparison with previous work that offers qualitative conclusions and comparisons among the study presented in this case-study

and the most recent studies tested on EMODB. Results indicate that the proposed approach outperforms several state-of-the-art approaches, achieving an accuracy of 87.7%.

The outline of the paper is as follows. Significant contributions to the research field of emotional recognition are presented in Sect. 2. The psychologically-inspired binary cascade classification schema is introduced and applied on EMODB in Sect. 3. In Sect. 4, the extracted features are demonstrated and in Sect. 5 the feature selection strategy is summarized. In Sect. 6, the experimental protocol is described. In the same section experimental results are presented. Some sets of alternative experiments are carried out and the corresponding results are discussed. Next, a statistical evaluation of the classifiers is carried out in order to verify though strong statistical tests which classifier is best performing for the task in question. A commentary to facilitate comparisons with most recent studies tested on the EMODB is provided in Sect. 7. Finally, conclusions are drawn in Sect. 8.

## 2 Related work

Emotions are fleeting, gender-dependent, and hard to distinguish. They exhibit substantial individual variations in expression and experience (Gunes et al. 2011). Emotion recognition is such a challenging task that it is unlikely to achieve perfect accuracy (Calvo and D'Mello 2011). Even humans may have difficulty describing how they feel, distinguishing between emotions or remembering how they felt only minutes earlier. A survey verifies the differences among human emotional perception (El Ayadi et al. 2011), whereas a more detailed study is available in (Dai et al. 2009). For the latter case, 20 subjects describe their perception of 6 emotions, namely happiness, hot anger, neutral, interest, panic, and sadness from the Emotional Prosody Speech and Transcripts (EPST) corpus. An accuracy of 64.4% is achieved by the humans with respect to the labels provided by the EPST corpus. Happiness/interest, happiness/neutral, interest/neutral, and neutral/sadness are the most confusing pairs.

Concerning the vocal part of affective computing, the most commonly used cues for emotion recognition are pitch and intensity. A survey (Zeng et al. 2007) verifies that speech is an important communication device in human communication. Authors claim that many studies indicate that the human judgement agreement is typically higher for facial expression modality than it is for vocal expression modality. The latter is verified in (Calvo and D'Mello 2011), where it is also stated that audio emotion recognition is low-cost, non-intrusive, and presents faster time resolution than facial emotion recognition.

Furthermore, it is underlined that the research community makes an effort to make use of contextual information, such as gender, to improve the performance of emotion recognition. The importance of context information is emphasized in (Calvo and D'Mello 2011). In an additional survey it is stated that besides features, also the applied classifier plays a significant role in emotion recognition performance (El Ayadi et al. 2011). For example, Gaussian mixture models (GMMs) cannot model the temporal structure of the data, whereas artificial neural networks classification accuracy seems to be fairly low when compared to other classifiers. On the contrary, SVMs appear to have global optimality of the training algorithm as well as high-performance data-dependent generalization bounds (El Ayadi et al. 2011).

Finally, although the vast majority of previous works exploit a speaker-dependent protocol, speaker-independent experiments are one of the latest trends in the emotion recognition field. Only a few researchers have conducted speaker-independent experiments to date. Here, we made a systematic effort to consider previous works that exploit the speaker-independent scenario, although they are quite sparse. By the term speaker-independent we mean that the utterances that are included in the test set come from one specific speaker, whose utterances are not included in the training set. In other words, it is not possible for the classifier to be tested on utterances derived from the same speaker whose utterances belong to the training set. In (El Ayadi et al. 2011), an alternative method for speaker-independency is proposed. In specific the authors mention that a speaker-independent emotion recognition system could be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system. However, the latter prerequisites that the speaker recognition system would ideally demonstrate an excellent performance.

### 2.1 Milestone emotion recognition systems

Lee and Narayanan (2005) present a case-study of detecting negative and non-negative emotions using spoken data from a call centre. The database is obtained from users engaged in spoken dialogue with a machine agent using a commercially-deployed call centre application. The authors apply linear discriminant classifiers (LDC) as well as $K$NN classifiers. The speech signal is analysed in fundamental frequency, energy, duration, and formant features. Then forward feature selection is applied followed by principal component analysis (PCA). Separate sets of experiments are carried out for male and female subjects. Results are reported for 10-fold cross-validation, making the experimental-procedure speaker-dependent. Concerning exclusively the audio channel, the lowest classification error is 17.85% for males and 12.04% for females, when LDC is applied.

Interest detection has been investigated in (Schuller et al. 2007, 2009a). In particular, three levels of interest are

identified: the first one includes disinterest, indifference, and neutrality, the second interest, and the third one curiosity. The authors record their own human conversations database, which they name AVIC database, with a duration of 10.5 hours. The following acoustic features are extracted: formants, pitch, frame energy, envelope, MFCC, harmonics-to-noise ratio, jitter, and shimmer. Then derivation of speed and acceleration regression coefficients of the aforementioned features takes place. Finally statistical functions are applied to the feature vector to render it less dependent of the spoken phonetic content. SVMs with a polynomial kernel are used as classifiers. Speaker-independent leave-one-speaker-out experiments are conducted. Feature selection is performed by sequential forward floating search at each iteration independently. A mean accuracy of 69.2% for all the three levels of interest is reported, when the audio channel is exploited exclusively.

### 2.2 Emotion recognition on EMODB

There is a number of recent contributions that implement emotion recognition on EMODB. In (Yang and Lugger 2010), the authors propose a set of harmony features. They are based on the psychoacoustic perception of pitch intervals and apply the theory of chords from music. Harmony features are derived from the pitch contour to characterize the relationship between different pitches, such as two-pitch intervals and chords involving more than two pitches. Harmony features are used in conjunction with energy, pitch, duration, formants, ZCR, and voice quality features. 306 statistical values of the aforementioned features are computed. Sequential floating forward selection identifies the 50 most informative features, which are fed as input to a Bayesian classifier that exploits GMMs. Speaker-independent experiments are carried out. 6 emotional classes are considered, namely: happiness, boredom, neutral, sadness, anger, and anxiety. In our approach a more exhaustive feature computation is available although in both works statistical values of features are computed and feature selection is applied in order to retain a small number of features. Nonetheless, in (Yang and Lugger 2010) disgust is dismissed.

The approach proposed in (Ruvolo et al. 2010) combines selection and hierarchical aggregation of features aiming to combine short, medium, and long time scale features. Considering short time scale features, MFCCs, sones, and linear predictive cepstral coefficients are used. Medium time scale features are computed by spectro-temporal box-filters, while long time scale features include phase, sampling interval, moment, energy, and summary statistics like mean value and quantiles. Next, GentleBoost is used to simultaneously select the best performing features and build the classifier. Speaker-independent experiments are performed. In specific, 63 binary classifiers are applied, each of which consists of 15 spectro-temporal box-filters selected by the GentleBoost. Finally, multinomial ridge logistic regression is applied to the continuous outputs of the 63 binary classifiers. The idea of calculating various features along with their corresponding statistics is also applied by the authors, although the categorization of features is not the same one. Moreover, the authors of this paper exploit feature selection and classification separately. Hierarchy is also applied in both approaches. However, in (Ruvolo et al. 2010) hierarchical aggregation of features is tested whereas in this approach the psychologically-inspired binary cascade classification schema employs a hierarchy on emotional descriptors.

Class-level spectral features for emotion recognition are proposed in (Bitouk et al. 2010). The authors define 3 phoneme type classes: stressed vowels, unstressed vowels, and consonants in the utterance. MFCC class-conditional means and standard deviations for each class are aggregated into one feature vector, by using the phoneme-level segmentation of the utterance. The average duration of the phoneme classes is appended to the feature vector. Moreover, 24 utterance level prosodic features are computed. The aforementioned features are related to statistics of fundamental frequency, first formant, voice intensity, jitter, shimmer, and relative duration of voiced segments. This results to a total of 261 features which are fed as input to a linear SVM classifier. A speaker-independent scenario is applied, whereas 6 emotional classes are taken into account, namely: anger, anxiety, disgust, happiness, sadness, and neutral. At a second set of experiments, feature selection is applied. Inspired by this approach, we performed sets of experiments with and without feature selection, as well. In the approach proposed by the authors of this paper, statistics of fundamental frequency, first formant, MFCCs, jitter, shimmer are among the computed features. However, in our case feature computation is exhaustive and it results a total number of 2327 extracted features.

An emotion recognizer that may operate jointly with an automatic speech recognizer is examined by (Pittermann et al. 2010). The feature vector comprises of MFCCs (along with their first- and second-order differences), intensity, and three formants, along with pitch and pitch statistics, namely minimum, mean, maximum, deviation and range. No feature selection technique is applied, while the HMMs are employed as classifiers to a speaker-dependent protocol, contrary to our approach that applies feature selection and a speaker-independent protocol. Also, speech recognition is not a prerequisite in this work, whereas the stated set of features in (Pittermann et al. 2010) is a subset of the feature vector computed by the authors. However, in both cases the authors compute the first- and second- differences of the features in order to capture their temporal evolution.

The aim of the work by (Altun and Polat 2009) is to improve the effectiveness of feature selection. 4 feature selection algorithms are examined, namely sequential forward selection, least square bound feature selection, mutual information based feature selection, and R2W2. 4 emotional states are taken into consideration, in specific, anger, happiness, neutral, and sadness. 58 features are extracted, 17 of which are prosodic, 5 correspond to sub-band energies, 20 MFCCs, and 16 LPCs. The emotion recognition problem is treated as a binary problem, which is also true for our approach. In particular, 2 frameworks are considered. The first framework is the "one-vs-rest" framework. In this approach the problem is to discriminate one emotional class from the rest (e.g. anger vs rest). The second framework is the "one-vs-one". In this case features that discriminate one emotional class from another one (e.g. anger vs happiness) are selected. For both frameworks, feature selection is applied for each sub-problem. 5 fold cross-validation is applied, rendering the experimental procedure speaker-dependent. SVMs with radial basis function kernels are utilized as classifiers. In this work, we resort to forward feature selection, since our aim is to improve classification accuracy, rather than exclusively the effectiveness of feature selection. It is true that we consider the additional emotional categories of disgust and anxiety. Accordingly, a different framework is applied, since each utterance of the test set may be classified to any of the seven emotional categories. However, SVM is among the classifiers examined by the authors of this paper.

A neutral utterance is considered as the reference and it is correlated with the rest of the utterances with means of a cross-correlogram (Chandaka et al. 2009). Five parameters are extracted as features from each cross-correlogram, named as peak value, instant at which peak occurs, centroid, equivalent width and mean square abscissa. Next, the parameters extracted from the cross-correlogram are fed as input to SVMs with Gaussian radial basis function kernel. Only the utterances derived from female speakers are considered. Speaker-dependent experiments are carried out, applying 50% of the utterances for training and the rest for testing, whereas no feature selection takes place, since the number of features is already restricted. Only 4 emotional states are considered, namely: anger, happiness, sadness, and neutral. In our work, we additionally consider boredom, disgust, and anxiety. Inspired by the features proposed in this paper, the authors decided to compute autocorrelation features as well as Teager energy operator on autocorrelation features along with their statistical values. A substantial difference of this work, is that for this work both male and female utterances are considered, either separately or combined. Additionally, the number of computed features by the authors of this paper is substantially greater and as a result feature selection is a prerequisite.

The importance of pitch contour is exhaustively examined in (Busso et al. 2009). Contrary to our work, the authors of (Busso et al. 2009) present a radically different approach, where just one feature category, namely pitch is sophisticatedly analysed. In a first step, reference pitch models for neutral speech are built. The neutral models are implemented by univariate GMMs. Then, the pitch-related input features are contrasted with the reference pitch models. In a second step, an LDC is implemented to classify the fitness measures. The fitness measures are the models of each pitch feature, that is represented with a univariate GMM. In essence, the fitness measure decides if the input speech is neutral or emotional speech, depending on whether it is similar to or different from the reference model, respectively. Speaker-dependent normalization, in specific pitch normalization and energy normalization, is used to reduce speaker variability. Two distinct sets of experiments are carried out: one at sentence level and another at voiced segments level. Different feature sets are selected for sentence and voiced segments level by exploiting forward feature selection in conjunction with logistic regression models. The sentence level accuracy equals 80.9%, whereas the corresponding accuracy for the voiced segments level is 71.1%, when all the seven emotional categories are considered. The authors agree with the just-described approach that normalization plays an essential role. The same is also true for feature selection. Finally, forward feature selection, is utilized in both papers.

Exploiting various feature categories is the aim of (Ramakrishnan and El Emary 2011). In specific, raw time signal, energy, spectral features, pitch, formants, MFCCs, cepstral features, voice quality, durational pause-related features, and Zipf features are fed as input to an SVM classifier. The recognition rate for anger is 0.96, for boredom 0.72, for disgust 0.71, for anxiety 0.81, for happiness 0.95, for neutral 0.68 and for sadness 0.75. As a next step individual features are tested for their recognition efficiency. Pitch, MFCCs, and formants are found to be those with the best discrimination ability. When confined to the aforementioned features, arousal equals 0.95 and valence 0.86, when SVM is employed as a classifier, whereas the corresponding rates diminish to 0.90 and 0.79, respectively for the case of HMM. In (Ramakrishnan and El Emary 2011) it is stated that the framework for speech emotion recognition comprises of feature extraction, feature selection, and classification. This approach is followed in the presented paper, as well. Also, the authors of this paper would like to verify that pitch, MFCCs, and formants seem to play a significant role for emotion recognition.

## 3 Psychologically-inspired binary cascade classification schema

In this section the EMODB database is briefly described (Burkhardt et al. 2005) and next the proposed psychologically-inspired binary cascade classification schema is presented. In EMODB 5 actors and another 5 actresses simulate 7 emotions. 10 utterances in German, which are used in everyday communication, are uttered by each actor/actress. 5 utterances are short, while the remaining 5 are long. The emotional labels are: anger, boredom, disgust, anxiety, happiness, sadness, and neutral. The audio format is PCM, the recordings are mono-channel, the sampling frequency is 16 KHz, and the audio samples are quantized in 16 bit. The full database comprises approximately 30 minutes of speech. To ensure emotional quality, a perception test was implemented. In specific, 20 subjects were asked to identify the emotional state of each utterance. Utterances whose emotion label is recognized correctly by at least 80% and possess naturalness higher than 60% are finally retained. This leads to a consistent database containing 535 utterances, 233 of which are uttered by male speakers, whereas the remaining 302 ones are uttered by female speakers. EMODB is annotated using ESPS/waves+. The main reason for choosing EMODB is the fact that it is consistently annotated and publicly available. Thus, it is a popular choice among researches, facilitating comparisons with previous work, which can be found in Sect. 7. Moreover, it covers most of the archetypal emotions (Ekman and Davidson 1994): happiness, sadness, fear, anger, surprise, and disgust. Archetypal emotions are the most long-standing way that emotion has been described by psychologists (Zeng et al. 2007). Finally, it is gender balanced.

Here, we propose a psychologically-inspired binary cascade classification schema that applies dimensional descriptions. Dimensional descriptions capture essential properties of emotional states. They are actually an alternative to the categorical description of human affect. In specific, dimensional descriptions characterize an emotional state in terms of a small number of latent dimension, whereas categorical descriptions consider a small number of discrete emotion categories. As stated in (Calvo and D'Mello 2011) it is important for computer scientists to utilize psychological emotion theories on their automatic emotion recognition systems. For this work we consider the dimensional descriptors (Ekman and Davidson 1994), making our approach psychologically founded. The first dimensional description is valence, also known as evaluation, a global measure of the pleasure associated with the emotional state, ranging from negative to positive. The second dimensional descriptor is activation, also known as arousal, that is how dynamic the emotional state is. Activation ranges from active to passive. The third dimensional descriptor is stance. Stance specifies how approachable, i.e. acceptable, the emotional state is. Positive values correspond to the advance approach whereas negative values correspond to the retreat approach.

The database is analysed into emotional categories with means of a binary tree structure, as is demonstrated in Fig. 1. The total number of nodes is $2 * 7 - 1 = 13$, where 7 is the number of emotional states included in the root of the tree. Firstly, a distinction between the non-negative valence and the negative valence emotions is made. Non-negative valence emotions include happiness and neutral, whereas negative valence emotions consist of anger, boredom, disgust, anxiety, and sadness. Secondly, a distinction between the non-negative valence feelings is carried out, separating happiness from the neutral state. Concerning negative valence emotions, a first grouping is done along the activation axis. Emotions with a negative activation are separated from those with a positive one. Boredom and sadness exhibit a negative activation, whereas anger, disgust, and anxiety present a positive activation. Then, boredom and sadness are separated. In the next step, a distinction among the positive activation emotions is carried out. Anger and anxiety belong to higher stance emotions and consequently they are grouped together, whereas disgust presents lower stance. In the final step, the higher stance emotions, i.e. anger and anxiety are separated.

Let us comment on the hierarchy of the emotional descriptors. Valence is chosen for the first level, since from a practical point of view it is fundamental to know whether the expressed emotion is negative or not. For example, it may be used to improve the quality of service in automated call centres. Furthermore, by discriminating negative from non-negative emotions, human-computer interaction designers will be able to recognize which parts of the interface are problematic, in the sense that they evoke negative emotions. Other possible applications include games, educational software, life- or in-car driver interfaces. At the second level, activation is applied. The reason for this choice is that when researchers limit themselves to a 2 dimensions emotional space instead of a 3 dimensions one, then this space is valence-activation (Bosma and André 2004; Cowie et al. 2001; Vogt et al. 2008). Obviously, to achieve better discrimination performance, through a finer filtering, stance is finally added to the dimensional descriptors.

The proposed psychologically-inspired binary cascade classification schema exhibits several advantages. To begin with, a feature that presents a good discrimination ability for a set of emotions may not present the same attribute for an alternative set of emotions. At the same time emotional states that share similar features can be grouped together. Thus, by confining the problem to a 2-class one, we manage to boost performance. For example, in (Fersini et al. 2009), the authors state that emotional categories anger/happiness, neutral/boredom, and

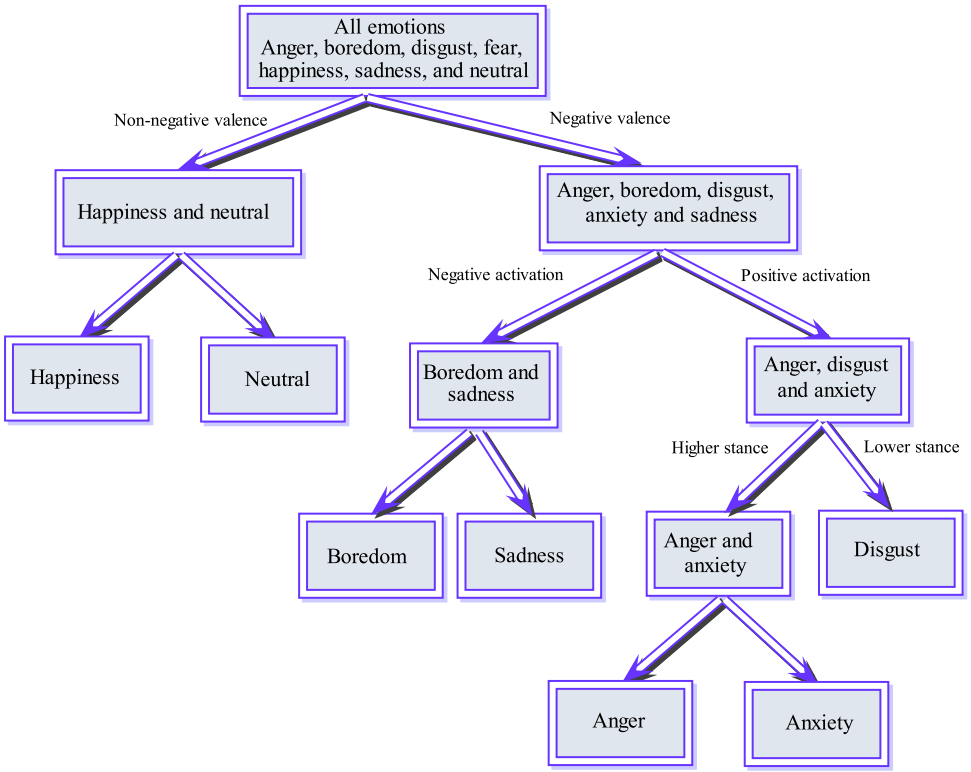


**Fig. 1** The proposed psychologically-inspired binary cascade classification schema for emotion recognition. Dimensional descriptions of valence, activation, and stance establish psychologically consistency

neutral/sadness are commonly confused. The latter is verified in (Ramakrishnan and El Emary 2011). However, with the proposed psychologically-inspired binary cascade classification schema the aforementioned couples of emotions belong to different analysis levels, rendering the discrimination between them more effective, as experimental results verify. Another advantage is that one can stop to whichever level is sufficient for the application. Thus, the proposed approach general and scalable. For example, detection of negative emotions, i.e. stopping to the 1st level, can be useful as a strategy to improve the quality of service in automated call centres. A further advantage of the proposed schema is that it can be easily adapted to additional problems that deal with different emotional labels by adding and/or removing emotional states, making the system a practical and flexible one. Finally, by considering the initial groups of emotions rather than emotions themselves, more utterances are available for the classifier to be trained, handling this way the common problem of the limited number of samples. It should also be mentioned that low-level attribute classifiers are more robust to data sparseness problem.

## 4 Feature extraction

In this section the extracted features studied in the proposed approach are outlined. Researchers have not yet found a sufficient feature set to reliably describe emotional states (Fersini et al. 2009; Zeng et al. 2007), which means that the problem of determination of the most informative features for emotion recognition an open one issue (Altun and Polat 2009). Emotion recognition performance can boost significantly if appropriate and reliable features are extracted (Altun and Polat 2009; Gunes et al. 2011; Yang and Lugger 2010). Highly informative features may be more critical to emotion recognition accuracy than the classifier itself (Altun and Polat 2009). To face this challenge in this case-study we accumulate a feature set as exhaustive as possible. Our aim is two-fold. On the one hand, we attempt to compute a multitude of features, aiming to capture as many informative features as possible. On the other hand, several features are investigated here for the first time for the emotion recognition task.

Several observations are made by the research community with respect to the correspondence among features and emotional categories. Previous work (Cowie et al. 2001) has indicated that anger presents an increased pitch, happiness demonstrates an increased intensity, whereas sadness exhibits a decrease in high-frequency energy. Sadness, anger, and fear exhibit the best recognition rates, whereas disgust has the worst (Calvo and D'Mello 2011). According to (Murray and Arnott 1993), for the feature of pitch range it is true that anger, happiness, and fear have a much wide pitch range, sadness has a slightly narrower, whereas disgust has a slightly wide one. Both pitch and energy for happiness and anger are usually higher than sadness (Iliou and Anagnos-

**Table 1** Formant contour related features

| Indices | Features |
|---|---|
| 1–4 | Mean of the 1st, 2nd, 3rd, and 4th formant |
| 5–8 | Maximum of the 1st, 2nd, 3rd, and 4th formant |
| 9–12 | Minimum of the 1st, 2nd, 3rd, and 4th formant |
| 13–16 | Variance of the 1st, 2nd, 3rd, and 4th formant |
| 17–20 | Skewness of the 1st, 2nd, 3rd, and 4th formant |
| 21–24 | Interquartile range of the 1st, 2nd, 3rd, and 4th formant |
| 25–28 | Range (i.e. maximum-minimum) of the 1st, 2nd, 3rd, and 4th formant |
| 29–32 | 90th percentile of the 1st, 2nd, 3rd, and 4th formant |

**Table 2** Pitch contour related features

| Indices | Features |
|---|---|
| 33–37 | Maximum, minimum, mean, median, interquartile range of pitch values (Sondhi 1968) |
| 38 | Pitch existence in the utterance expressed in percentage (0–100%) (Sondhi 1968) |
| 39–42 | Maximum, mean, median, interquartile range of durations for the plateaux at minima (Sondhi 1968) |
| 43–45 | Mean, median, interquartile range of pitch values for the plateaux at minima (Sondhi 1968) |
| 46–50 | Maximum, mean, median, interquartile range, upper limit (90%) of durations for the plateaux at maxima (Sondhi 1968) |
| 51–53 | Mean, median, interquartile range of pitch values within the plateaux at maxima (Sondhi 1968) |
| 54–57 | Maximum, mean, median, interquartile range of durations of the rising slopes of pitch contours (Sondhi 1968) |
| 58–60 | Mean, median, interquartile range of pitch values within the rising slopes of pitch contours (Sondhi 1968) |
| 61–64 | Maximum, mean, median, interquartile range of durations of the falling slopes of pitch contours (Sondhi 1968) |
| 65–67 | Mean, median, interquartile range of pitch values within the falling slopes of pitch contours (Sondhi 1968) |
| 68–75 | Maximum, minimum, mean, variance, median, skewness, interquartile range, and 90th percentile of pitch (Boersma 1993) |
| 76–83 | Maximum, minimum, mean, variance, median, skewness, interquartile range, and 90th percentile of first-order differences of pitch (Boersma 1993) |
| 84–91 | Maximum, minimum, mean, variance, median, skewness, interquartile range, and 90th percentile of second-order differences of pitch (Boersma 1993) |
| 92 | Number of frames, where pitch value could not be determined (Boersma 1993) |

topoulos 2009). Fear and anger present a high energy lever, whereas sadness demonstrates a low one (Scherer 2003). For anger and happiness the pitch contour is higher and more variable than for sadness (Juslin and Laukka 2003). Previous work (Bosma and André 2004), has indicated that the acoustic features related to pitch, energy, and speaking rate are not appropriate to model valence. For example there is a tendency to confuse hot anger to happiness and interest to sadness. According to (Ramakrishnan and El Emary 2011) fear resembles sadness having an almost downwards slope in the pitch contour, whereas joy exhibits a rising slope. Anger, fear, happiness, and surprise appear to share the same characteristics with respect to the fundamental frequency (El Ayadi et al. 2011). Pitch range, mean of fundamental frequency, mean intensity, speech rate, and high-frequency energy appears to be an index into arousal. Shorter pauses and inter-breath stretches are indicative of higher activation. Fast speaking rate, less high-frequency energy, low pitch and large pitch range and longer vowel durations are related to positive valence (Gunes et al. 2011).

We extract features related to statistics of formant contours (Table 1), pitch contours (Table 2), and energy contours (Table 3). The method to estimate formants relies on the linear prediction analysis (Markel and Gray 1976), whereas pitch is computed based on an autocorrelation method. Statistics related to the distribution of energy into several spectral bands (Table 4) as well as statistics of the TEO-autocorrelation (Table 5) are also computed. The latter is used to detect creaky voice. Furthermore, features stemming from generic audio classification are extracted, as is summarized in Table 6. Such features include spectral, temporal, perceptual, short-time energy, and MPEG-7 standard descriptors (Benetos et al. 2007). Moreover, motivated by the high speech emotion recognition accuracy reported in (Zervas et al. 2006), when Fujisaki's model parameters (Mixdorff 2000) are considered, these features are also tested here, as can be seen in Table 7. In addition, jitter and shimmer, are listed in Table 8. Statistics of the contours of zero-crossing rate (ZCR), autocorrelation, and Spectrum Flux are summarized in Table 9. Linear prediction filter coefficients (LPC)-related features are exhibited in Table 10, whereas perceptual linear predictive (PLP)- and relative spectral perceptual linear predictive (RASTA-PLP)-related ones in Table 11. For Tables 5–11 with the term statistics we refer to maximum, minimum, variance, mean, median, skewness, interquartile range, and 90th percentile) of the respective feature with this order. Statistics have proven to be less sensitive to linguistic information (Ramakrishnan and El Emary 2011).

First and second order differences are computed to capture the features temporal evolution. Thus, a multivariate time series is obtained and a dynamic model is feasible (Schuller et al. 2007). Statistical functions are used to summarize features. Moreover, they reduce the dependency on the spoken phonetic content (Schuller et al. 2007, 2009a). This way emotion recognition accuracy is improved (Ruvolo et al. 2010). In particular, the statistical functions which are computed to those features that this is applicable are:

**Table 3** Energy contour related features

| Indices | Features |
|---|---|
| 93–97 | Maximum, minimum, mean, median, interquartile range of energy values (Ververidis and Kotropoulos 2006) |
| 98–101 | Maximum, mean, median, interquartile range of durations for the plateaux at minima (Ververidis and Kotropoulos 2006) |
| 102–104 | Mean, median, interquartile range of energy values for the plateaux at minima (Ververidis and Kotropoulos 2006) |
| 105–109 | Maximum, mean, median, interquartile range, upper limit (90%) of duration for the plateaux at maxima (Ververidis and Kotropoulos 2006) |
| 110–112 | Mean, median, interquartile range of energy values within the plateaux at maxima (Ververidis and Kotropoulos 2006) |
| 113–116 | Maximum, mean, median, interquartile range of durations of the rising slopes of energy contours (Ververidis and Kotropoulos 2006) |
| 117–119 | Mean, median, interquartile range of energy values within the rising slopes of energy contours (Ververidis and Kotropoulos 2006) |
| 120–123 | Maximum, mean, median, interquartile range of durations of the falling slopes of energy contours (Ververidis and Kotropoulos 2006) |
| 124–126 | Mean, median, interquartile range of energy values within the falling slopes of energy contours (Ververidis and Kotropoulos 2006) |
| **127–134** | Maximum, minimum, variance, mean, median, skewness, interquartile range, and 90th percentile of first-order differences of energy values (Kotti and Kotropoulos 2008) |
| **135–142** | Maximum, minimum, variance, mean, median, skewness, interquartile range, and 90th percentile of second-order differences of energy values (Kotti and Kotropoulos 2008) |
| 143–144 | Position of the first energy maximum/minimum (Kotti and Kotropoulos 2008) |
| 145–149 | Maximum, minimum, mean, median, variance of the temporal distance between two successive energy maxima (Kotti and Kotropoulos 2008) |
| 150–154 | Maximum, minimum, mean, median, variance of the temporal distance between two successive energy minima (Kotti and Kotropoulos 2008) |
| 155–159 | Maximum, minimum, mean, median, variance of the temporal distance between two successive energy extrema (i.e. either maximum/minimum or minimum/maximum) (Kotti and Kotropoulos 2008) |
| **160–161** | Standard deviation of energy rising/falling energy slopes (Kotti and Kotropoulos 2008) |

maximum, minimum, variance, mean, median, skewness, interquartile range, and 90th percentile.

The pitch contour is one of the important properties of speech that is affected by the emotional modulation (Fersini et al. 2009; Iliou and Anagnostopoulos 2009; Lee and Narayanan 2005; Pantic and Rothkrantz 2003; Vanello et al. 2009), since emotional pitch modulation is triggered by the activation level of the sentence (Busso et al. 2009). Previous work (Busso et al. 2009) has indicated that features such as mean, maximum, minimum, and range describe the global aspects of pitch contour and are more emotionally salient than features that describe the pitch shape itself, like slope, curvature, and inflexion. Energy, formants, voice quality, and spectral features efficiency is already verified (Dai et al. 2009; Fersini et al. 2009; Iliou and Anagnostopoulos 2009; Inanoglu and Caneel 2005; Kostoulas and Fakotakis 2006; Lee and Narayanan 2005; Schuller et al. 2008). Less frequently used are the LPCs, PLPs, and RASTA-PLPs (Pao et al. 2006; Zervas et al. 2006). MFCC effectiveness in proven in (Mishra et al. 2009; Sato and Obuchi 2007). Jitter is also investigated in (Vanello et al. 2009). Shimmer is also investigated in (Espinosa and Reyes-García 2009). MPEG-7 descriptors, TEO-autocorrelation, total loudness, and SLSC are investigated in this study for the first time within the context of emotion recognition to the best of the authors' knowledge. That comprises a set of 602 features. The corresponding indexes are: {127–142, 160, 161, 190–213, 262–549, 1126–1397} and are noted in bold in Tables 1–11.

**Table 4** Features related to the distribution of energy in several frequency bands (Ververidis and Kotropoulos 2006)

| Indices | Features |
|---|---|
| 162–169 | Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, 3950 Hz |
| 170–176 | Energy in the frequency bands 250–600, 600–1000, 1000–1500, 1500–2100, 2100–2800, 2800–3500, 3500–3950 Hz |
| 177–182 | Energy in the frequency bands 250–1000, 600–1500, 1000–2100, 1500–2800, 2100–3500, 2800–3950 Hz |
| 183–187 | Energy in the frequency bands 250–1500, 600–2100, 1000–2800, 1500–3500, 2100–3950 Hz |
| 188–189 | Energy ratio between the frequency bands (3950–2100) and (2100–0) and between the frequency bands (2100–1000) and (1000–0) |

**Table 5** Statistics of TEO-autocorrelation features (Kotti and Kotropoulos 2008)

| Indices | Parameter |
|---|---|
| **190–197** | TEO-autocorrelation |
| **198–205** | First-order differences of TEO-autocorrelation |
| **206–213** | Second-order differences of TEO-autocorrelation |

## 5 Selecting small feature sets

This Section focuses on selecting a small set of features that would be late on fed as input to classifiers. However, before applying feature selection it is important to pre-process our

**Table 6** Statistics of generic audio classification features (Benetos et al. 2007)

| Indices | Parameter |
|---|---|
| 214–221 | Short-term energy |
| 222–229 | Short-term energy first-order differences |
| 230–237 | Short-term energy second-order differences |
| 238–245 | Audio fundamental frequency |
| 246–253 | Audio fundamental frequency first-order differences |
| 254–261 | Audio fundamental frequency second-order differences |
| **262–269** | Total loudness |
| **270–277** | Total loudness first-order differences |
| **278–285** | Total loudness second-order differences |
| **286–349** | First 8 specific loudness sensation coefficients |
| **350–413** | First 8 specific loudness sensation coefficients first-order differences |
| **414–477** | First 8 specific loudness sensation coefficients second-order differences |
| **478–485** | AudioSpectrumCentroid |
| **486–493** | AudioSpectrumCentroid first-order differences |
| **494–501** | AudioSpectrumCentroid second-order differences |
| **502–509** | AudioSpectrumRolloff frequency |
| **510–517** | AudioSpectrumRolloff frequency first-order differences |
| **518–525** | AudioSpectrumRolloff frequency second-order differences |
| **526–533** | AudioSpectrumSpread |
| **534–541** | AudioSpectrumSpread first-order differences |
| **542–549** | AudioSpectrumSpread second-order differences |
| 550–741 | 24-order MFCCs |
| 742–933 | 24-order MFCCs first-order differences |
| 934–1125 | 24-order MFCCs second-order differences |
| **1126–1138** | 13 autocorrelation coefficients |
| **1139** | Log-attack time |
| **1140** | Temporal centroid |
| **1141–1172** | AudioSpectrumFlatness (4 coefficients) |
| **1173–1204** | AudioSpectrumFlatness first-order differences |
| **1205–1236** | AudioSpectrumFlatness second-order differences |

features carefully in order to guarantee the quality of data. For that reason feature removal and feature normalization are carried out.

### 5.1 Feature removal

Certain features have to be removed, because many missing values are observed. For example, there are specific pitch contours that do not have plateaux below 45% of their maximum pitch value whereas other features do not have first- and/or second-order differences. When more than 2% of the total number of feature values is missing, the corresponding feature is discarded. For the EMODB, the indices of the discarded features are {145–159, 243, 250, 251, 258, 259, 1238, 1280, 1326, 1361, 1363, 1364, 1369, 1371, 1372, 1374, 1377, 1379, 1380, 1385, 1387, 1388, 1393, 1395, 1396, 1468, 1491}. That is, finally, the remaining features are $2327 - 41 = 2286$.

**Table 7** Statistics of Fujisaki's model parameters (Mixdorff 2000)

| Indices | Parameter |
|---|---|
| **1237–1244** | Fujisaki's $F0$ contour |
| **1245–1252** | Fujisaki's $F0$ contour first-order derivative |
| **1253–1260** | Fujisaki's $F0$ contour second-order derivative |
| **1261–1268** | Fujisaki's logarithmic $F0$ spline |
| **1269–1276** | Fujisaki's logarithmic $F0$ spline first-order derivative |
| **1277–1284** | Fujisaki's logarithmic $F0$ spline second-order derivative |
| **1285–1292** | Low-pass filter output contour |
| **1293–1300** | High-pass filter output contour |
| **1301–1308** | Accent component |
| **1309–1316** | Accent component first-order derivative |
| **1317–1324** | Accent component second-order derivative |
| **1325–1332** | Phrase component |
| **1333–1340** | Phrase component first-order derivative |
| **1341–1348** | Phrase component second-order derivative |
| **1349–1356** | Accent commands |
| **1357–1364** | Accent commands first-order derivative |
| **1365–1372** | Accent commands second-order derivative |
| **1373–1380** | Phrase commands |
| **1381–1388** | Phrase commands first-order derivative |
| **1389–1396** | Phrase commands second-order derivative |
| **1397** | Base frequency (single value parameter) |

### 5.2 Feature normalization

As a prerequisite to feature selection, normalization takes place. Feature normalization improves the features generalization ability (Bosma and André 2004) and guarantees that all the features obtain the same scale (Lee and Narayanan 2005) in order to ensure an equal contribution of each feature to the feature selection algorithm (Ververidis and Kotropoulos 2005). Moreover, normalization helps to address the problem of outliers. All features are subject to min-max normalization hereafter. Min-max normalization is expected to boost performance, since it preserves all original feature relationships and does not introduce any bias in the features.

### 5.3 Feature selection

Feature selection presents several advantages. To begin with, small feature subsets require less memory and computations, whereas they also allow for a more accurate statis-

**Table 8** Statistics of jitter, shimmer, and voice quality (Boersma 1993)

| Indices | Parameter |
|---|---|
| 1398 | Number of pulses |
| 1399 | Number of periods |
| 1400 | Mean period |
| 1401 | Standard deviation of period |
| 1402 | Fraction of locally unvoiced frames |
| 1403 | Number of voice breaks |
| 1404 | Degree of voice breaks |
| 1405 | Jitter local |
| 1406 | Jitter absolute |
| 1407 | Jitter relative average perturbation |
| 1408 | Jitter five-point period perturbation quotient |
| 1409 | Jitter average absolute difference between consecutive difference between consecutive periods divided by the average period |
| 1410 | Shimmer local |
| 1411 | Shimmer average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20 |
| 1412 | Shimmer three-point amplitude perturbation quotient |
| 1413 | Shimmer five-point amplitude perturbation quotient |
| 1414 | Shimmer eleven-point amplitude perturbation quotient |
| 1415 | Shimmer average absolute difference between consecutive difference between consecutive periods |
| 1416 | Mean autocorrelation |
| 1417 | Mean noise to harmonics ratio |
| 1418 | Mean harmonics to noise ratio |

**Table 9** Statistics of ZCR, autocorrelation, and Spectrum Flux (Benetos and Kotropoulos 2010)

| Indices | Parameter |
|---|---|
| 1419–1426 | ZCR |
| 1427–1434 | First-order differences of ZCR |
| 1435–1442 | Second-order differences of ZCR |
| 1443–1450 | autocorrelation (8 coefficients) |
| 1451–1458 | First-order differences of autocorrelation |
| 1459–1466 | Second-order differences of autocorrelation |
| 1467–1474 | Spectrum Flux |
| 1475–1482 | First-order differences of Spectrum Flux |
| 1483–1490 | Second-order differences of Spectrum Flux |

tical modeling, thus improving performance. On the contrary, large feature sets may yield a prohibitive computational time for classifier training. For example, neural networks face difficulties, when they are fed with extraordinary many features. Employing large feature sets increases the risk of including features with reduced discriminating power. Additionally, feature selection eliminates irrelevant

**Table 10** The first 16 LPCs, their first- and second-order differences, along with their statistics (Jackson 1989)

| Indices | Parameter |
|---|---|
| 1491–1506 | LPCs |
| 1507–1521 | First-order differences of LPCs |
| 1522–1535 | Second-order differences of LPCs |
| 1536–1543 | Statistics of LPCs |
| 1544–1551 | Statistics of first-order differences of LPCs |
| 1552–1559 | Statistics of second-order differences of LPCs |

**Table 11** Statistics of 16-order PLPs and 16-order RASTA-PLPs (Ellis 2005)

| Indices | Parameter |
|---|---|
| 1560–1687 | PLPs |
| 1688–1815 | PLPs first-order differences |
| 1816–1943 | PLPs second-order differences |
| 1944–2071 | RASTA-PLPs |
| 2072–2199 | RASTA-PLPs first-order differences |
| 2200–2327 | RASTA-PLPs second-order differences |

features, leading to a reduction in the cost of acquisition of the data. Furthermore, if all the features are employed, there is the curse of dimensionality as well as the risk for overfitting. In addition, feature selection can boost performance when the number of training utterances is not sufficient or when a real-time problem needs to be handled.

Context is a major factor in affective computing (Picard 1997), since the interpretation of human behavioral signals is context dependent (Zeng et al. 2007). Modern psychology suggests that the emotion should be described in terms of appraisal dimensions which give information about the context in which the emotion is produced. Consequently, an important related issue that should be addressed in emotion recognition is how one can make use of information about the context of the emotional behavior. Here, we consider gender as one context parameter. It is widely accepted by the research community that gender normalization should be applied (Busso et al. 2009), since the two genders convey their emotions in profoundly different ways (Zeng et al. 2007). For that reason many contributions consider each gender separately (Chandaka et al. 2009; Ververidis and Kotropoulos 2005). The reason is that some features are gender-dependent (Pantic and Rothkrantz 2003), e.g. the pitch (Pittermann et al. 2010). It is well known that in general female speech has higher pitch than male speech, due to the increase in mass of a male's vocal folds. Furthermore, smaller vocal tract dimensions in women rather than in men leads to the production of higher formant frequencies for women. Another example is the speaking rate of males that is lower than

that of females while expressing anger or disgust (Ramakrishnan and El Emary 2011). So, in this case-study, feature selection is applied separately for male and for female subjects.

Feature selection is applied at every level of the psychologically-inspired binary cascade classification schema, in order to allow refinement for the specific level of the schema and consequently to boost performance. That is, different features are expected to distinguish negative from non-negative valence emotions to those which distinguish between happiness and neutral state. Also, a specific feature **x** that presents a good discrimination ability between a couple of emotional states set, may not bear the same attribute for an alternative set.

Forward Feature Selection (FFS) strategy is used in this case study which is simple, fast, effective and widely accepted technique (Guyon and Elisseeff 2003). Two classes are considered in each step, according to the proposed psychologically-inspired binary cascade classification schema. For example, when the emotions of negative valence are separated from those with non-negative valence, the one class consists of the features derived from utterances labeled as happiness and neutral and the second one of those labeled as anger, boredom, disgust, anxiety, and sadness.

Here, a total of 75 features are selected for each emotion recognition sub-problem. We concluded to the aforementioned number of features after experimental evaluation. In specific, several numbers of features ranging from 50 to 100 have been tested with respect to emotion recognition accuracy of the first level of the psychologically-inspired binary cascade classification schema, that is the recognition accuracy in discriminating negative from non-negative valence emotions. We decided to confine ourselves to a small number of features, since the more complicated the classifier schema along with an extended feature set creates concerns about overfitting. Moreover, a small feature set renders the system a fast, thus a practical one. 75 features were found to have the best performance (emotion recognition accuracy equal to 97.0%) for the linear SVM. This procure has been verified to be successful in (Schuller et al. 2005b). For reasons of homogeneity the same number of features is retained for the remaining levels of the psychologically-inspired binary cascade classification schema. We resorted to linear SVM in this phase, because it needs no parametrization and also because SVMs are ideal for binary classification problems. The percentage distribution of selected feature sets along the emotional groups, as they appear in Fig. 1, can be seen in Table 12 for each gender separately.

## 6 Classification

In this section the speaker-independent experimental protocol is initially presented (Sect. 6.1). Next, classification results using either $K$NNs (Sect. 6.2) or SVMs (Sect. 6.3) are computed. For the SVM case two kernels are tested, namely the Gaussian radial basis function and the linear kernel. In Sect. 6.4 some sets of additional experiments are carried out in order to quantify the efficiency improvement achieved by the proposed contributions. Finally, strong statistical analysis of the results takes place in Sect. 6.5.

### 6.1 Speaker-independent experimental protocol

The experimental protocol chosen includes conduction of speaker-independent experiments. In order to measure speaker-independent emotion recognition rate, leave-one-speaker-out evaluation is applied. That is, for each gender separately, the classifier is trained 5 times, each time leaving one speaker out of the training set and then testing the performance on the left out speaker. Final results are reported along the two genders.

Speaker-independent systems present several advantages. They are able to handle efficiently an unknown speaker. Thus, they are more robust and stable, and demonstrate a better generalization ability than the speaker-dependent ones, since they avoid classifier overfitting. In speaker-dependent experimentation it is possible that the classifier may learn special characteristics of the specific speaker along with the emotional attributes whereas in speaker-independent experiments the classifier is forced to learn exclusively the emotional attributes, decoupling in this way emotion recognition from speaker recognition. As discussed in the following, this is the reason why speaker-dependent experiments are expected to demonstrate better emotion recognition accuracy when compared to speaker-independent ones. In addition, speaker-independent models are useful when there is not enough data to build a sufficient and reliable model or when the training procedure is not completed (Schuller et al. 2005a). Furthermore, speaker-independent systems can cope with the limited number of speakers. An additional advantage of the speaker-independency is the fact that the experimental protocol is deterministic, in the sense that the exact configuration is known. On the contrary when speaker-dependent cross validation is employed, the random partition does not allow for a reproduction of the exact configuration, making the results not directly comparable between research groups (Schuller et al. 2009b). Finally, the speaker-independent systems are suitable for real-life applications, such as call centre applications, media segmentation, and public transport surveillance.

As already noted, speaker-dependent emotion recognition leads to far better results than speaker-independent modeling. Previous work (Austermann et al. 2005) has indicated that an average emotion recognition rate of 84% is achieved in speaker-dependent experiments, whereas for

**Table 12** Selected feature set (%) per gender and emotion group

| | {happiness, neutral} vs {anger, boredom, disgust, anxiety, sadness} | | {happiness} vs {neutral} | | {boredom, sadness} vs {anger, disgust, anxiety} | | {boredom} vs {sadness} | | {anger, anxiety} vs {disgust} | | {anger} vs {anxiety} | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Formant contour | 2.7 | 1.4 | 1.4 | 5.3 | 1.3 | 1.3 | 1.4 | 1.3 | 1.3 | 2.7 | 1.3 | 1.3 |
| Pitch contour | 2.7 | 1.4 | 6.7 | 2.7 | 2.7 | 5.3 | 2.7 | 1.3 | 2.7 | 5.3 | 4.0 | 2.7 |
| Energy contour | 0 | 4.0 | 0 | 2.7 | 0 | 4.0 | 2.7 | 0 | 4.0 | 0 | 1.3 | 2.7 |
| Energy distribution | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.3 | 1.3 | 2.7 | 1.3 | 1.3 |
| TEO-autocorrelation | 0 | 0 | 1.4 | 0 | 0 | 0 | 1.4 | 0 | 2.7 | 1.3 | 1.3 | 1.3 |
| Generic audio classification | 52 | 58.4 | 45.3 | 50.7 | 42.7 | 26.7 | 46.6 | 53.3 | 50.7 | 42.7 | 58.7 | 44.0 |
| Fujisaki model parameters | 4.0 | 2.7 | 9.3 | 4.0 | 2.7 | 9.3 | 4.2 | 4.0 | 4.0 | 4.0 | 5.3 | 0 |
| Jitter & shimmer | 1.4 | 0 | 1.4 | 0 | 1.4 | 4.0 | 2.7 | 0 | 0 | 1.3 | 0 | 1.3 |
| ZCR, autocorrelation, Spectrum Flux | 1.4 | 4.0 | 2.6 | 2.6 | 5.3 | 8 | 2.7 | 0 | 2.7 | 0 | 2.7 | 4.0 |
| LPCs | 1.4 | 1.4 | 2.6 | 8 | 6.6 | 10.7 | 0 | 2.8 | 0 | 1.3 | 1.3 | 1.3 |
| PLPs | 34.7 | 26.7 | 29.3 | 24 | 37.3 | 30.7 | 35.6 | 36.0 | 30.6 | 38.7 | 22.8 | 40.1 |

the speaker-independent case the emotion recognition drops to 60%. The aforementioned conclusion is also verified in (Schuller et al. 2005b) for 10 different classifiers. The averaged emotion recognition performance equals 89.49% for the speaker-dependent case, whereas it drops to 71.29% for the speaker-independent one. In addition, activation classification to 3 categories, namely high, neutral, and low, is carried out in (Tato et al. 2002). Speaker-independent average activation recognition equals 59.3%, whereas when speaker-dependent experiments are executed the average activation recognition raises to 83.7%. This conclusion is also verified in (Bitouk et al. 2010), where speaker-independent and speaker-dependent experiments lead to an emotion recognition rate equal to 78.2% for the first case and to 84.8% for the second case.

Efficiency is presented by means of confusion matrices as well as emotion recognition accuracy for this case-study. The columns of the confusion matrix correspond to the actual emotion and the rows to the predicted one. The confusion matrix elements are recognition rates expressed in percentage. Confusion matrices presented in Tables 13–15 are column normalized confusion matrices, since the number of utterances per each emotional class is not the same. Accuracy is the percentage of the correctly recognized utterances to the total number of utterances.

### 6.2 $K$ nearest neighbors ($K$NNs)

The $K$NN classifier is used as a baseline classifier. If $K = 1$, all utterances derived from the training set will be classified correctly, but the test set accuracy will be insufficient. As $K \to \infty$, a less biased classifier is obtained. However, since the number of utterances is a finite one, it is true that optimality cannot be reached. In this paper, $K$NN with Euclidean distance function is tested. A varying number of neighbors is investigated ranging from 1 to 20. The way emotion recognition accuracy develops with respect to parametrization is demonstrated in Fig. 2. To comment on the just-mentioned figure, the male curve corresponds to the mean male accuracy over 5 experiments. As a reminder, there are 5 male speakers in EMODB and the experiments are speaker-independent. This means that the classifier is trained using 4 male speakers and tested on the remaining 5th one. The aforementioned protocol is repeated 5 times, so that every male speaker appears in the test set exactly once. Next, the mean accuracy over the five male subjects is reported. The same protocol is also applied for the 5 female speakers. The term "all subjects", stands for the averaged mean accuracy over male and female subjects together.

In general, emotion recognition accuracy of female subjects tends to be higher than that of male subjects for low values of $K$, whereas it inclines to be lower than that of male subjects for high $K$ values. Overall emotion recognition accuracy improves continuously from $K = 1$ to $K = 6$, where it reaches its maximum value. The confusion matrix for $K = 6$ is provided in Table 13 and the corresponding emotion recognition accuracy succeeded is 77.9%. For $K > 6$ values the emotion recognition accuracy presents a generally falling trend with a slow rate. Thus, the best parameter is indicated by experimentation, an experimental protocol proven to be successful in (Burkhardt et al. 2006; Chandaka et al. 2009; Fersini et al. 2009; Lee and Narayanan

**Table 13** Confusion matrix (%) for the $K$NN with 6 neighbors

| | Happiness | Neutral | Boredom | Sadness | Disgust | Anxiety | Anger |
|---|---|---|---|---|---|---|---|
| Happiness | 75.4 | 1.8 | 0 | 0 | 0 | 0 | 1.7 |
| Neutral | 16.0 | 76.4 | 1.3 | 0 | 0 | 0 | 0 |
| Boredom | 1.4 | 19.1 | 89.3 | 13.7 | 4.9 | 0 | 0.8 |
| Sadness | 0 | 2.7 | 8.1 | 86.3 | 0 | 10.0 | 0 |
| Disgust | 2.9 | 0 | 1.3 | 0 | 48.8 | 15.7 | 4.9 |
| Anxiety | 2.9 | 0 | 0 | 0 | 29.2 | 64.3 | 4.9 |
| Anger | 1.4 | 0 | 0 | 0 | 17.1 | 10.0 | 87.7 |

**Table 14** Confusion matrix (%) for the SVM with Gaussian radial basis function kernel ($\sigma = 1$)

| | Happiness | Neutral | Boredom | Sadness | Disgust | Anxiety | Anger |
|---|---|---|---|---|---|---|---|
| Happiness | 89.7 | 0 | 0 | 0 | 0 | 0 | 1.7 |
| Neutral | 2.9 | 90.5 | 1.5 | 0 | 0 | 0 | 0 |
| Boredom | 0 | 9.5 | 91.0 | 11.4 | 2.5 | 0 | 0 |
| Sadness | 0 | 0 | 6.0 | 88.6 | 0 | 3.1 | 0 |
| Disgust | 2.9 | 0 | 1.5 | 0 | 47.5 | 6.1 | 1.7 |
| Anxiety | 0 | 0 | 0 | 0 | 37.5 | 87.7 | 6.5 |
| Anger | 4.5 | 0 | 0 | 0 | 12.5 | 3.1 | 90.1 |

**Table 15** Confusion matrix (%) for the linear SVM

| | Happiness | Neutral | Boredom | Sadness | Disgust | Anxiety | Anger |
|---|---|---|---|---|---|---|---|
| Happiness | 92.6 | 0 | 0 | 0 | 0 | 0 | 1.6 |
| Neutral | 0 | 98.9 | 1.3 | 0 | 0 | 0 | 0 |
| Boredom | 0 | 1.1 | 88.5 | 11.5 | 2.5 | 0 | 0 |
| Sadness | 0 | 0 | 8.9 | 88.5 | 0 | 3.1 | 0.8 |
| Disgust | 2.9 | 0 | 1.3 | 0 | 60.0 | 9.2 | 3.3 |
| Anxiety | 2.9 | 0 | 0 | 0 | 27.5 | 83.1 | 6.6 |
| Anger | 1.6 | 0 | 0 | 0 | 10.0 | 4.6 | 87.7 |

2005; Schuller et al. 2007, 2009a; Ververidis and Kotropoulos 2005). If parameters are randomly or badly chosen, the worst emotion recognition accuracy equals about 70%.

### 6.3 Support vector machines (SVMs)

SVMs are also known as maximum margin classifiers. The SVM theory is formulated to solve binary classification problems originally, rendering SVMs ideal for the case under consideration, since we apply a psychologically-inspired binary cascade classification schema. Two different kernels are exploited. Let $\mathbf{v}_i$ be the $i$th training vector.

1. Gaussian radial basis function kernel:

$$K_{SVM}(\mathbf{v}_i, \mathbf{v}_j) = \exp\left(-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|^2}{2\sigma^2}\right), \quad (1)$$

   where $\sigma$ is a scaling factor; and
2. Linear (homogeneous):

$$K_{SVM}(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{v}_i^T \mathbf{v}_j. \quad (2)$$

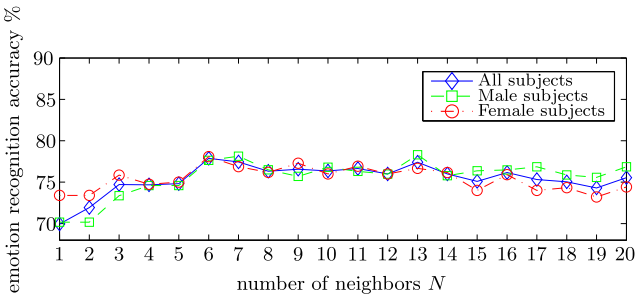


**Fig. 2** Speaker-independent emotion recognition accuracy of $K$NN for various $K$ values, for male subjects, female subjects, and both genders

SVM with Gaussian radial basis function kernel are tested for various $\sigma$ values with $\sigma \in (0, 10]$, as can be seen in Fig. 3. The best performance is obtained for $\sigma = 1$. For the case of SVM with Gaussian radial basis function kernel the two genders exhibit the same pattern: emotion recognition accuracy reaches at a fast rate its maximum for $\sigma = 1$,

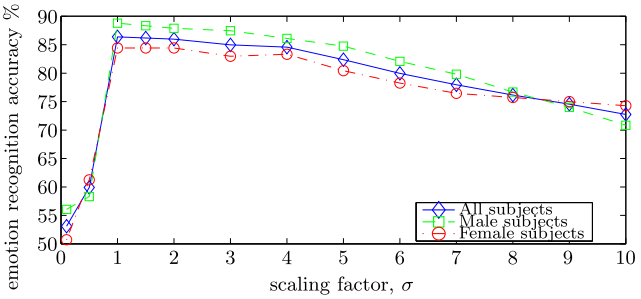


**Fig. 3** Speaker-independent emotion recognition accuracy of SVM with Gaussian radial basis function kernel for various $\sigma$ values, for male subjects, female subjects, and both genders

whereas it decreases strictly at a slower rate for greater $\sigma$ values. The confusion matrix for $\sigma = 1$ is exhibited in Table 14 and the related accuracy equals 86.4%. Male emotion recognition accuracy is consistently greater than female emotion recognition, with exception of extreme low and high $\sigma$ values. The lower bound accuracy presented by SVM with Gaussian radial basis function kernel is 50.7% and can be attributed to poor parametrization. Linear SVM has the advantage of no need for parametrization. The corresponding confusion matrix is sketched in Table 15. It achieves an emotion recognition accuracy equal to 87.7%, which is in fact the highest achieved by all three classifiers.

### 6.4 Discussion

Before proceeding to the statistical analysis of the experimental results, we would like to discuss some issues related to the performance of the proposed approach.

With respect to the first contribution of the paper, i.e. the psychologically-inspired binary cascade classification schema, an additional set of experiments has taken place in order to verify its effectiveness. For the aforementioned experiments we resorted to a flat, one-layer, single multiclass system (i.e. the classifier discriminates among the seven emotions in just one step) instead of the binary cascade classification schema. The rest of the experimental protocol is identical to that of the proposed approach. Once again, 75 features are retained from a total of 2286 features for each gender separately. However, this time feature selection takes place just once and not for every single step of the psychologically-inspired binary cascade classification schema. In total, 75 features are selected for the male subjects and 75 for the female ones. This means that the 75 selected features are supposed to be discriminative for a 7-class problem, instead of a 2-class one. An emotion recognition accuracy of 65.2% is reported, when the $K$NN with 6 neighbours is exploited. $K$NN is preferred since SVM is inherently a two-class classifier.

Referring to the second contribution, that is the use of the 602 novel features, a second set of alternative experiments has been carried out. In this case, the novel features are retained from the original feature set, whereas the remaining experimental protocol remains unaltered. That is, the psychologically-inspired binary cascade classification schema is exploited for each gender separately, but there are no novel features among the 75 selected ones. If the reader inspects Tables 1–11, he/she has to ignore those feature whose indices are in bold. In essence, the total number of computed features is 1725, minus those that are removed because of the missing values, that is a total of 1703 features prior to feature selection. In that case an overall emotion classification accuracy of 72.4% is reported, when SVM with linear kernel is applied as classifier. With respect to features computation, an extra set of experiments has been executed. If all 2286 features are fed as input to the linear SVM classifier, i.e. no feature selection is applied prior to classification, the performance achieved equals 66.5%. Once again, the remaining experimental procedure parameters are not modified. The aforementioned fact underlines the importance of feature selection, whereas the interested reader may find a discussion on the subject on Sect. 5.3.

Regarding the last contribution, that is the separation of genders, a fourth set of complementary experiments has been completed. In this case we retain the psychologically-inspired binary cascade classification schema, as well as the full feature set, comprising of all 2286 features before feature selection and of 75 features after feature selection. This time, however, the features are selected for both genders simultaneously, as traversing the classification schema. In specific, feature selection is carried out exactly half times compared to those required for the original proposed system. For this set of experiments, however, the 75 selected features try to discriminate between two emotional categories for male and female subjects, simultaneously. Moreover, with respect to the speaker-independent protocol the classifier is trained on 9 subjects and tested on the refrained one. This means that both males as well as females subjects are included in the training procedure, regardless of the testing speaker. Specifically, for 5 experiments the classifier is trained on 4 male and 5 female subjects and tested on restrained 1 male subject, whereas for additional 5 experiments, 5 male and 4 female subjects are exploited to train the classifier, whereas the testing is carried out with the remaining 1 female subject. When the linear SVM is tested as classifier, the reported emotion classification accuracy equals 72.2%.

A positive point of the work presented here is that it manages to separate among negative emotions effectively. However, previous research has shown that distinguishing among negative emotions is a more difficult task than distinguishing among positive ones (Watson 2000). Furthermore, the presented strategy discriminates efficiently between anger and happiness, a problem commonly met in the literature (Chandaka et al. 2009; Yang and Lugger 2010).

With respect to our previous work (Kotti et al. 2010), a smaller set of 1418 features was computed to discriminate

between negative and non-negative valence emotions. The latter is just the first level of the proposed psychologically-inspired binary cascade classification schema presented in Fig. 1. The best accuracy achieved in (Kotti et al. 2010) equals about 90.0%. Here with the extraction of 2327 features the corresponding accuracy (i.e. the emotion recognition accuracy between negative and non-negative valence emotions) raises up to 97.0%, verifying the discriminative power of the features added to the feature set, as well as the importance of the computation of adequate features for the emotion recognition task.

It is worth to note that here the linear SVM has accomplished emotion recognition accuracy equal to 87.7% for all 7 emotions and 97.0% for discriminating negative from non-negative valence emotions. The complete set of 7 emotions corresponds to the lowest binary tree nodes, whereas the couple of negative and non-negative valence emotions is, in essence, the first level of the psychologically-inspired binary cascade classification schema. The performance decrease may be attributed to a couple of facts. On the one hand, further misclassifications may happen as the binary tree is traversed, while on the other hand less utterances are available for each refinement step.

### 6.5 Statistical analysis

In this section, two sets of comparisons are carried out. The first one refers to studying the classifiers' error rate differences, so as to test whether the classifiers are statistically significantly different or not. The second one assesses the information expressed by the confusion matrices and ranks the classifiers according to their degree of informativeness. Statistical analysis of results is a desirable procedure, since it adds to the meta-knowledge of the experimental outcome. For example, it provides insight about the classifiers suitability, strengths, flexibility, and effectiveness for the problem in question. This way, the best performing solutions are highlighted and proposed for future exploration based on a strong mathematical background. Moreover, statistical analysis may be a prerequisite step for creating an ensemble of classifiers.

With respect to the first comparison set, we would like to examine the accuracy of the three different classifiers. To do so, the method described in (Guyon et al. 1998) is applied, which exploits the Normal law distribution to determine if the error rate difference between a couple of classifiers is statistically significant or not. It is, in essence, a hypothesis testing, where under hypothesis $H_0$ the error rates of the classifiers are considered to be equal, whereas under the alternative hypothesis $H_1$ the first classifier is better than the second one. Since it is a statistical method, a confidence interval has to be determined. Here, the confidence interval is set to 95%. For the $K$NN and the SVM with Gaussian radial basis function kernel the best parametrization case is considered, i.e. $K = 6$ and $\sigma = 1$, respectively. First, we compare the SVM with Gaussian radial basis function kernel against the $K$NN and we find that the SVM with Gaussian radial basis function kernel is statistically significant better than the $K$NN. Next, the comparison of the linear SVM classifier against the $K$NN takes place and the $K$NN is found to be statistically significant worst that the linear SVM. Finally we compare the linear SVM against the SVM with Gaussian radial basis function kernel. This time the classifiers are found to be of equal accuracy, consequently is evaluated that the SVM kernel does not play a statistically important role in the classifier accuracy.

Moving to the second set of comparisons, we aim to assess the classifiers performance in terms of confusion matrices evaluation (MacKay 2003). In (Wallach 2006), it is demonstrated that for comparing confusion matrices the common metrics of accuracy, precision, recall, and $F_1$ may be inappropriate, since they can be misleading about evaluating the information expressed by the confusion matrix. On the contrary mutual, information is considered to be the appropriate measure to compare classifiers' confusion matrices. For this purpose, the mutual information between the predicted label and the ground truth label is computed for each confusion matrix (Wallach 2006). For the case of the $K$NN classifier, the mutual information equals 0.5095, for the SVM with Gaussian radial basis function kernel the corresponding value is 0.5524, whereas for the linear SVM the mutual information value raises to 0.5609. This means that the most informative classifier is the linear SVM, less informative is the SVM with Gaussian radial basis function kernel, and the least informative is the baseline $K$NN classifier.

## 7 Comparison with previous related work utilizing EMODB

As stated in Sect. 2.2, there is a number of works that implement emotion recognition on EMODB. The aim of this Section is to present, discuss and compare the contemporary results presented in the aforementioned recent approaches to those of ours approach.

In (Yang and Lugger 2010), the rates of the correctly classified utterances are as follows: 52.7% for the happiness, 84.8% for boredom, 52.9% for neutral, 87.6% for sadness, 86.1% for anger, and 76.9% for anxiety. The authors come to the conclusion that no feature set can sufficiently distinguish between anger and happiness. In specific, 33.9% of the utterances that express happiness are wrongly classified as expressing anger. However, the strategy proposed here is able to distinguish between anger and happiness. As is demonstrated in Table 15, 1.6% of the angry utterances

are mistakenly classified as utterances expressing happiness, whereas 1.6% of the happy coloured utterances are erroneously classified as angry ones. Furthermore, the authors state that boredom and anger are easily separated, which is also verified by the experiments presented in this paper (see Tables 13–15).

The overall emotion recognition accuracy achieved in (Ruvolo et al. 2010) equals 78.7% for all seven emotional states of EMODB. In specific, anger is recognized correctly at a 92.91% rate, boredom at 74.68%, disgust at 68.42%, anxiety at 70.91%, happiness at 50%, neutral at 85.90%, and sadness at 90.57%. Compared to the approach presented in this paper, the greatest difference in recognition rate refers to happiness, which we classify correctly with a 92.6% recognition rate (absolute improvement equal to 42.6 points), while the lower difference is observed for the utterances expressing sadness, which are classified in our work at a 87.7% recognition rate (absolute deterioration equal to 5.21 points).

An emotion recognition rate of 78.2% is achieved, without feature selection in (Bitouk et al. 2010). With feature selection, emotion recognition accuracy equals 78.5% for rank search SVM wrapper, 81.3% for rank search subset evaluation, 78.2% for greedy stepwise SVM wrapper, and 79.1% for information gain ratio. For the best performing case, that is rank search subset evaluation, the technique proposed in this paper presents an accuracy improvement of 6.4 points.

Emotion accuracy equals 71.7% for all emotional classes in (Pittermann et al. 2010), that is an absolute deterioration of 16 points when compared to the system presented here. The authors in (Pittermann et al. 2010) pay attention to the frequently confused pairs of emotions. According to reported results, anger and happiness exhibit the maximum average error frequency equal to 6.66%, while in our approach the corresponding rate is 1.6%. The second most commonly confused pair of emotions is fear and happiness. For the method presented in (Pittermann et al. 2010), 6.04% of the utterances expressing anxiety are wrongly classified as happiness, while for our approach the corresponding rate is 2.9%. The least confusing pairs of emotion in (Pittermann et al. 2010) are anger and sadness, as well as fear and sadness that are never confused, which is also true for our approach. Finally, when emotion recognition is coupled with the recognizer output voting error reduction, which is a speech recognition technique, emotion recognition accuracy raises up to 76.4%.

The best accuracy for the 4 emotional classes of anger, happiness, neutral, and sadness equals 85.5% in (Altun and Polat 2009). It is achieved when the least square bound feature selection algorithm is applied and 13 features are retained. Our best accuracy equals 87.7% for all 7 emotional classes. Confining ourselves to the emotional categories recognized in (Altun and Polat 2009), i.e. anger, happiness, neutral, and sadness, for our approach the emotional classes of happiness, neutral, and sadness are those with the 3 highest recognition rates (sadness and boredom share the same recognition rate of 88.5%), while anger has the 4th highest recognition rate.

The overall accuracy achieved reaches 84.55%. The percentage of the correctly classified utterances equals 95.04% for anger, 66.67% for happiness, 84.34% for sadness, and 85.07% for neutral. In (Chandaka et al. 2009), as in the work presented in (Yang and Lugger 2010), there is a high confusion between happiness and anger, which is not the case of our study. More specifically, in (Chandaka et al. 2009), 18.05% of the utterances that express actually happiness are classified as conveying anger. It should be noted that with respect to happiness our algorithm manages an absolute maximum improvement of 25.93 points. On the contrary, the greatest (and only) deterioration of our algorithm, compared to the one presented in (Chandaka et al. 2009), refers to the emotional category of anger and equals 7.34 points. Finally, both the approach presented in (Chandaka et al. 2009) as well as our approach manage a perfect separation between neutral and sadness utterances.

## 8 Concluding remarks

This paper addresses the vocal emotion recognition problem by investigating a large feature set and a psychologically-inspired binary cascade classification schema. The proposed psychologically founded schema applies the "divide-and-conquer" technique, discriminating initially among emotion categories rather than emotions themselves. This way, commonly confused pairs of emotions are clearly separable. The schema is easily adaptable to diverse emotions and its analysis level is adjustable to the problem under consideration.

There are three main steps in the presented approach: first feature extraction is applied, then feature selection is carried out, and finally classification is employed. So far the research community has not agreed on a feature set that efficiently describes the emotional states. The authors' contribution to investigate the aforementioned matter lies in the computation of an extended feature set. Several features that are considered here, are proposed for the first time within the context of emotion recognition. The emotional categories of the features are related to statistics of pitch, formants, and energy contours, as well as spectrum, cepstrum, perceptual and temporal features, autocorrelation, MPEG-7 descriptors, Fujisaki's model parameters, voice quality, jitter, and shimmer. Feature selection is executed by means of FFS for each gender separately, aiming to take context into account. EMODB is utilized to conduct speaker-independent experiments. The aforesaid means that the same speaker may not occur in the training and the test splits, making

the system able to handle efficiently an unknown speaker. Therefore, there is no risk of classifier overfitting, as well as speaker adaptation. Consequently, the proposed system is robust, stable, and is expected to demonstrate a generalization ability. The baseline $K$NN, as well as SVMs with linear and Gaussian radial basis function kernel are examined as classifiers.

Finally, statistical analysis of the experimental results is carried out with respect to emotion recognition accuracy as well as in terms of confusion matrices evaluation. Analysis verifies that SVMs are suitable classifiers for emotion recognition under the binary cascade classification schema. In particular, linear SVM, which has the advantage of no parametrization, accomplishes a high accuracy of 87.7%, better than the accuracy presented in recent work exploiting EMODB.

Future work includes the computation of more features, aiming to capture additional aspects of emotions. Furthermore, alternative feature selection techniques may be tested and compared. Additionally, classification results could be fused by means of a parallel or a tandem Bayesian network, so as to lead to a more robust system and to boost performance further. Towards the same goal meta-classifiers could be applied, as well. It would be also interesting to ask human listeners to annotate the database and then compare the classifier confusion matrices to the human ones. Moreover, an additional classification schema could be applied that exploits exclusively the dimensional descriptors, i.e. valence, activation, and stance, instead of the categorical ones.

**Acknowledgements** M. Kotti would like to thank Associate Professor Constantine Kotropoulos for his valuable contributions for the extraction of part of the features that are described in Sect. 4.

## References


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.

Austermann, A., Esau, N., Kleinjohann, L., & Kleinjohann, B. (2005). Prosody based emotion recognition for MEXI. In *Proc. IEEE/RSJ int. conf. intelligent robots and systems*, Edmonton, Canada, August 2005 (pp. 201–208).

Benetos, E., & Kotropoulos, C. (2010). Non-negative tensor factorization applied to music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1955–1967.

Benetos, E., Kotti, M., & Kotropoulos, C. (2007). Large scale musical instrument identification. In *Proc. 4th sound and music computing conference*, Lefkada, Greece, July 2007 (pp. 283–286).

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*(7–8), 613–625.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. institute of phonetic sciences* (Vol. 17, pp. 97–110).

Bosma, W., & André, E. (2004). Exploiting emotions to disambiguate dialogue acts. In *Proc. 9th int. conf. intelligent user interfaces*, Funchal, Portugal, January 2004 (pp. 85–92).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proc. 9th European conf. speech communication and technology*, Lisbon, Portugal, September 2005 (pp. 1517–1520).

Burkhardt, F., Ajmera, J., Englert, R., Stegmann, J., & Burleson, W. (2006). Detecting anger in automated voice portal dialogs. In *Proc. 9th int. conf. spoken language processing*, Pittsburgh, USA, September 2006 (pp. 1–4).

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Speech and Audio Processing*, *17*(4), 582–596.

Calvo, R. A., & D'Mello, S. (2011). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37.

Chandaka, S., Chatterjee, A., & Munshi, S. (2009). Support vector machines employing cross-correlation for emotional speech recognition. *Measurement*, *42*(4), 611–618.

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

Dai, K., Fell, H., & MacAuslan, J. (2009). Comparing emotions using acoustics and human perceptual dimensions. In *Proc. 27th int. conf. extended abstracts on human factors in computing systems*, Boston, USA, April 2009 (pp. 3341–3346).

Ekman, P., & Davidson, R. (1994). *The nature of emotion: fundamental questions*. New York: Oxford University Press.

Ekman, P., Matsumoto, D., & Friesen, W. (2005). Facial expression in affective disorders. In *Series in affective science. What the face reveals* (pp. 331–342). London: Oxford Press. Chap. 15.

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.

Ellis, D. P. W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. URL http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/. Online web resource.

Espinosa, H. P., & Reyes-García, C. (2009). Detection of negative emotional state in speech with anfis and genetic algorithms. In *Proc. 6th int. workshop models and analysis of vocal emissions for biomedical applications*, Florence, Italy, December 2009 (pp. 24–28).

Fersini, E., Messina, E., Arosio, G., & Archetti, F. (2009). Audio-based emotion recognition in judicial domain: A multilayer support vector machines approach. In *Proc. 6th int. conf. machine learning and data mining in pattern recognition*, Leipzig, Germany, July 2009 (pp. 594–602).

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of IEEE int. conf. automatic face and gesture recognition*, Santa Barbara, USA, March 2011 (pp. 827–834).

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(7–8), 1157–1182.

Guyon, I., Makhoul, J., Schwartz, R., & Vapnik, V. (1998). What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(1), 52–64.

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., & Friedman, S. (2005). Distinguishing deceptive from non-deceptive speech. In *Proc. 9th European conf. speech communication and technology*, Lisbon, Portugal, September 2005 (pp. 1833–1836).

Iliou, T., & Anagnostopoulos, C. (2009). Statistical evaluation of speech features for emotion recognition. In *Proc. 4th int. conf. digital telecommunications*, Colmar, France, July 2009 (pp. 121–126).

Inanoglu, Z., & Caneel, R. (2005). Emotive alert: HMM-based emotion detection in voicemail messages. In *Proc. 10th int. conf. intelligent user interfaces*, San Diego, USA, January 2005 (pp. 251–253).

Jackson, L. B. (1989). *Digital filters and signal processing* (2nd ed.). New York: Kluwer Academic.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814.

Konstantinidis, E. I., Hitoglou-Antoniadou, M., Luneski, A., Bamidis, P. D., & Nikolaidou, M. M. (2009). Using affective avatars and rich multimedia content for education of children with autism. In *Proc. 2nd int. conf. pervsive technologies related to assistive environments*, Corfu, Greece, June 2009 (pp. 1–6).

Kostoulas, T. P., & Fakotakis, N. (2006). A speaker dependent emotion recognition framework. In *Proc. 5th int. symposium communication systems, networks and digital signal processing*, Patras, Greece, July 2006 (pp. 305–309).

Kotti, M., & Kotropoulos, C. (2008). Gender classification in two emotional speech databases. In *Proc. 19th int. conf. pattern recognition*, Tampa, USA, December 2008 (pp. 1–4).

Kotti, M., Paternò, F., & Kotropoulos, C. (2010). Speaker-independent negative emotion recognition. In *Proc. 2nd int. workshop cognitive information processing*, Elba Island, Italy, June 2010.

Lee, C. M., & Narayanan, S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(12), 293–303.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech*. New York: Springer.

Minker, W., Pittermann, J., Pittermann, A., Strauß, P. M., & Bühler, D. (2007). Challenges in speech-based human–computer interfaces. *International Journal of Speech Technology*, *10*(2–3), 109–119.

Mishra, H. K., & Sekhar, C. C. (2009). Variational Gaussian mixture models for speech emotion recognition. In *Proc. 7th int. conf. advances in pattern recognition*, Kolkata, India, February 2009 (pp. 183–186).

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proc. IEEE int. conf. acoustics, speech, and signal processing*, June 2000 (pp. 1281–1284).

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097–1108.

Nass, C., Jonsson, I. M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In *Proc. int. conf. human-computer interaction, extended abstracts on human factors in computing systems*, Portland, OR, USA, April 2005 (pp. 1973–1976).

Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2009). An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech, and Music Processing*. doi:10.1155/2009/594103.

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390.

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2006). Human computing and machine understanding of human behavior: A survey. In *Proc. 8th int. conf. multimodal interfaces*, Banff, Canada, November 2006 (pp. 239–248).

Pao, T. L., Chen, Y. T., Yeh, J. H., & Li, P. J. (2006). Mandarin emotional speech recognition based on SVM and NN. In *Proc. 18th int. conf. pattern recognition*, Hong Kong, Hong Kong, August 2006 (pp. 1096–1100).

Picard, R. W. (1997). *Affective computing*. Cambridge: MIT Press.

Pittermann, J., Pittermann, A., & Minker, W. (2010). Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, *13*(1), 49–60.

Ramakrishnan, S., & El Emary, I. (2011). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 1–12. doi:10.1007/s11235-011-9624-z.

Ruvolo, P., Fasel, I., & Movellan, J. R. (2010). A learning approach to hierarchical feature selection and aggregation for audio classification. *Pattern Recognition Letters*, *31*(12), 1535–1542.

Sato, N., & Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Journal of Natural Language Processing*, *14*(4), 83–96.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005a). Speaker independent speech emotion recognition by ensemble classification. In *Proc. IEEE int. conf. multimedia and expo*, Amsterdam, The Netherlands, July 2005 (pp. 864–867).

Schuller, B., Villar, R. J., Rigoll, G., & Lang, M. (2005b). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In *Proc. IEEE int. conf. acoustics, speech, and signal processing*, Philadelphia, USA, March 2005 (pp. 325–328).

Schuller, B., Müeller, R., Höernler, B., Höethker, A., Konosu, H., & Rigoll, G. (2007). Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of 9th int. conf. multimodal interfaces*, Nagoya, Japan, November 2007 (pp. 30–37).

Schuller, B., Rigoll, G., Can, S., & Feussner, H. (2008). Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In *Proc. 17th IEEE int. symposium robot and human interactive communication*, Munich, Germany, August 2008 (pp. 453–458).

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., & Konosu, H. (2009a). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, *27*(12), 1760–1774.

Schuller, B., Steidl, S., & Batliner, A. (2009b). The INTERSPEECH 2009 emotion challenge. In *Proc. 10th annual int. conf. speech communication association*, Brighton, UK, September 2009 (pp. 312–315).

Sondhi, M. M. (1968). New methods of pitch extraction. *IEEE Transactions on Audio and Electroacoustics*, *16*(2), 262–266.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). Emotional space improves emotion recognition. In *Proc. 7th int. conf. spoken language processing*, September 2002 (pp. 2029–2032).

Vanello, N., Martini, N., Milanesi, M., Keiser, H., Calisti, M., Bocchi, L., Manfredi, C., & Landini, L. (2009). Evaluation of a pitch estimation algorithm for speech emotion recognition. In *Proc. 6th int. workshop models and analysis of vocal emissions for biomedical applications*, Florence, Italy, December 2009 (pp. 29–32).

Ververidis, D., & Kotropoulos, C. (2005). Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm. In *Proceedings of IEEE int. conf. multimedia and expo*, Los Alamitos, USA, July 2005 (pp. 1500–1503).

Ververidis, D., & Kotropoulos, C. (2006). Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In *Proc. 14th European signal processing conference*, Florence, Italy, September 2006.

Vogt, T., André, E., & Bee, N. (2008). EmoVoice—A framework for online recognition of emotions from voice. In *Proc. 4th IEEE tutorial and research workshop on perception and interactive technologies for speech-based systems*, Irsee, Germany, June 2008 (pp. 188–199).

Wallach, H. (2006). *Evaluation metrics for hard classifiers* (Tech. Rep.). Cambridge University, Cavendish Lab. URL www.inference.phy.cam.ac.uk/hmw26/papers/evaluation.ps.

Watson, D. (2000). *Mood and temperament*. New York: Guilford Press.

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing, Special Section on Statistical Signal & Array Processing*, *90*(5), 1415–1423.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2007). A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *Proc. 9th int. conf. multimodal interfaces*, Nagoya, Japan, November 2007 (pp. 126–133).

Zervas, P., Mporas, I., Fakotakis, N., & Kokkinakis, G. K. (2006). Employing Fujisaki's intonation model parameters for emotion recognition. In *Proc. 4th hellenic conf. artificial intelligence*, Heraclion, Greece, May 2006 (pp. 443–453).