EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Perceptual audio features for emotion detection

Mehmet Cenk Sezgin, Bilge Gunsel* and Gunes Karabulut Kurt

Abstract

In this article, we propose a new set of acoustic features for automatic emotion recognition from audio. The features are based on the perceptual quality metrics that are given in perceptual evaluation of audio quality known as ITU BS.1387 recommendation. Starting from the outer and middle ear models of the auditory system, we base our features on the masked perceptual loudness which defines relatively objective criteria for emotion detection. The features computed in critical bands based on the reference concept include the partial loudness of the emotional difference, emotional difference-to-perceptual mask ratio, measures of alterations of temporal envelopes, measures of harmonics of the emotional difference, the occurrence probability of emotional blocks, and perceptual bandwidth. A soft-majority voting decision rule that strengthens the conventional majority voting is proposed to assess the classifier outputs. Compared to the state-of-the-art systems including Munich Open-Source Emotion and Affect Recognition Toolkit, Hidden Markov Toolkit, and Generalized Discriminant Analysis, it is shown that the emotion recognition rates are improved between 7-16% for EMO-DB and 7-11% in VAM for "all" and "valence" tasks.

Keywords: perceptual audio feature extraction, audio emotion recognition, PEAQ

1. Introduction

It is well known that human speech contains not only the linguistic content, but also the emotion of the speaker. The emotion may play a key role in many applications like in entertainment electronics to gather emotional user behaviors, in Automatic Speech Recognition to resolve "how it was said" other than "what it was said", and in text-to-speech systems to synthesize emotionally more natural speech. Therefore, in human-machine interaction applications, it is important that emotional states in human speech are fully perceived by computers [1,2].

However, detecting the emotion content of an audio signal has several challenges. The main difficulties stem from the fact that it is quite difficult to define what emotion means and how it can be categorized in a precise way [3]. There are ongoing debates concerning how many emotion categories exist, whether the categories should classically be represented in a discrete (i.e., sad, happy) or continuous manner and how to approach long-term and short-term transitions of emotional states and debates as to how to seek measurable correlates of emotions. Therefore, different approaches exist to model emotions in the psychological literature [4], details of which will be provided in Section 6. In this study, we use two-dimensional continuous space, which is commonly used with the purpose of benchmarking among various emotion corpora which are already mapped onto diverse emotion groups [5].

Another critical research challenge in the emotion-detection problem is to determine the features that influence the recognition of emotion in speech [6]. The precise feature extraction from subjective patterns such as emotion is a highly challenging issue and depends strongly on the application and database at hand. There is considerable uncertainty as to the best feature set for classifying emotional data and which classifiers to use. The existence of different contents, genders, speakers, and speaking styles raise complications because these properties have direct affect on the features such as pitch, and energy contours [2].

Existing emotion-detection methods make use of acoustic features which are mostly related to speech recognition; fundamental frequency or pitch, energy, speaking rate, and spectral coefficients such as Mel-frequency cepstral coefficients (MFCCs) [2,6]. There are two standard popular and freely available toolkits using these generic feature sets: the Munich Open-Source

* Correspondence: gunselb@itu.edu.tr
Multimedia Signal Processing and Pattern Recognition Lab., Department of Electronics and Communications, Istanbul Technical University, Istanbul, Turkey

Emotion and Affect Recognition Toolkit (openEAR) and the Hidden Markov Toolkit (HTK). HTK [7] is a very basic approach employing the features that is used for a very broad selection of speech and general audio-recognition tasks. The openEAR is a more specifically tailored emotion-detection tool [8], which extracts more than 4,000 features by 39 functionals of 56 acoustic low-level descriptors (LLDs) and corresponding first- and second-order delta regression coefficients. Recently, a Generalized Discriminant Analysis (GerDa) method is proposed based on deep neural networks [9]. GerDa is able to learn 2D features extracted from 6552-dimensional openEAR features. However, it requires a semi-supervised pre-optimization of the several free parameters corresponding to hidden layers of the GerDa deep neural network.

Since the conventional features are originally proposed for speech recognition, they may not fully model the emotion perception of the human ear [10,11]. This is because a vast majority of the traditional features, such as MFCC, are generated for short speech frames to decode the phonemes. However, the emotional state of the speaker is unlikely to change as fast as phonemes [11,12]. Consequently, until now, a high performance emotion detector could only be achieved by using very large feature sets [5,7] or small feature sets in combination with highly complex classifiers [9]. The results in the literature verify that existing features are not adequate especially in the valence domain where dimensions categorize emotions according to them being pleasant or unpleasant.

In this article, we propose a set of acoustic features which are designed to detect the perceptual content of the speech rather than the conventional features that are targeted for speech recognition. Thus, we define a new method which is more compliant with the subjective nature of the human emotion depending on the perceptual evaluation of human auditory system. The proposed set referred as *perceptual feature* set consists of a 9-dimensional feature vector with 7 low level and 2 statistical descriptors. We use support vector machine (SVM) and Gaussian mixture model (GMM) as the primary classifiers which efficiently model diverse statistics of the emotional data. Preliminary version of the perceptual feature set is presented as a conference paper in [13]; however, in this study, a new decision rule that improves the recognition rate is designed and used for the emotion recognition. The new decision rule referred as soft-majority voting (S-MV) forces the classification as a combination of minimum variance with the highest posterior probabilities for each category thus strengthens the majority voting.

We benchmarked our emotion classification performance with the performance reported in [5,9,14], which is achieved by HTK [7], openEAR [8], and GerDa [9] tools, for comparison purposes. We report the results on two databases, i.e., EMO-DB [15] and VAM [16]. Our test results demonstrate that on the average, the *perceptual feature* set outperforms the legacy features in terms of classification accuracy for valence. The rest of this article is organized as follows. The related study is summarized in Section 2. The emotional datasets used in tests are described in Sections 3. The concept of emotional variances and the mathematical derivation of the perceptual features are, respectively, introduced in Sections 4 and 5. The test results are summarized in Sections 6. Finally, conclusions are given in Section 7.

2. Related study

In this section, we provide an overview on the conventional features and the emotional category definitions in the literature. We also point out the problems encountered in existing emotion-recognition systems. Our initial consideration is the emotional categorization methods where the interpretation accuracy of expressions and physiological responses is challenging. Later, we will look into the conventional features employed in emotion detection and will discuss the associated problems.

2.1. Emotion representation in discrete and continuous spaces

According to research in psychology, three major approaches are of concern that affect emotion modeling: *categorical*, *dimensional*, and *appraisal-based* approach. Since the *appraisal-based* approach is not prevalently used because of its complex and sophisticated measurements of change [4], we concentrate on the mostly employed categories; the *categorical* and the *dimensional* approaches.

Categorical approach considers the definition of diverse emotion classes that are basic and popular universally, called basic emotions. Six basic emotions are defined by Ekman [17] which we are familiar with; happiness, sadness, anger, fear, surprise, and disgust. However, people may reveal rather complex emotional modes; therefore, a single label or discrete class may not reflect the actual affective state [4].

An alternative methodology is the utilization of continuous emotion dimensions. The use of *dimensional* description of human affect defines the dependency of the categories of one another; rather than their dependency as in categorical description. A three-*dimensional* emotion space is proposed: arousal (activation), potency (power), and valence (pleasure) evaluation [18]. Another alternative is simpler two-*dimensional* emotion space: arousal and valence. Yet, the most widely used *dimensional* model is based on the assumption of Russell [19] that each basic emotion is represented by a bipolar entity being a part of the same emotional dimension in two-dimensional emotion space. The proposed poles are relaxed versus aroused for the arousal and pleasant versus unpleasant for the valence. We can conclude that as the *categorical* approach discretizes the emotion space model into classical fragments, the *dimensional* method defines a continuous emotion space which is accepted more associated with the real life experiences. According to these findings, we have used *dimensional* approach as will be further described in Section 6.

2.2. Conventional features in audio emotion detection

A proper selection of features plays a substantial role in the classification performance, since pattern recognition techniques are generally dependent on the problem domain. Therefore, selection of the suitable features is an important issue in the design of an audio emotion detection system.

Acoustical speech features reported in the literature are shown in Figure 1[2]. Existing systems use a number of integrated continuous, qualitative, spectral as well as the Teager energy operator (TEO)-based features. Energy and pitch are the primary continuous acoustical features which are heavily used in emotion recognition. Since arousal refers to the amount of energy required to express a certain emotion, according to the studies reported in [2], the arousal state of the speaker affects the overall energy, energy distribution across the frequency spectrum, and the frequency and duration of pauses of speech signal. On the other hand, since the short-term speech energy is closely related with activation or arousal dimension of the emotion, its usage in the conventional features contributes to the classification of emotions which have similar arousal level [20].

Other continuous features are related to the fundamental frequency (*F0*), the articulation rate, and the spectral information in voiced and unvoiced portions of speech. The fundamental frequency that is produced by the pitch signal, also known as the glottal waveform, carries emotional information because of its dependency on the tension of the vocal folds and the subglottal air pressure. The vibration of the vocal folds is the source of the pitch signal. The time elapsed between two successive vocal fold openings is called pitch period *T*, while the vibration rate of the vocal folds is the fundamental frequency of the phonation *F0* or pitch frequency [6]. High glottal volume velocity indicates a music like speech like joy or surprise and low velocity stands for modes such as anger or disgust.

Numerous features are applied to describe the shape of the vocal tract which is modified by the emotional modes. Formants are one of the leading features which represent vocal tract resonances that form the spectrum [6]. On the other hand, the emotional content of an utterance is strongly related to its voice quality [1,2,6]. According to Cowie et al. [1], the voice quality features are grouped into four measurement categories: (i) voice level, (ii) voice pitch, (iii) phrase, phoneme, word, and feature boundaries; (iv) temporal structures. However,

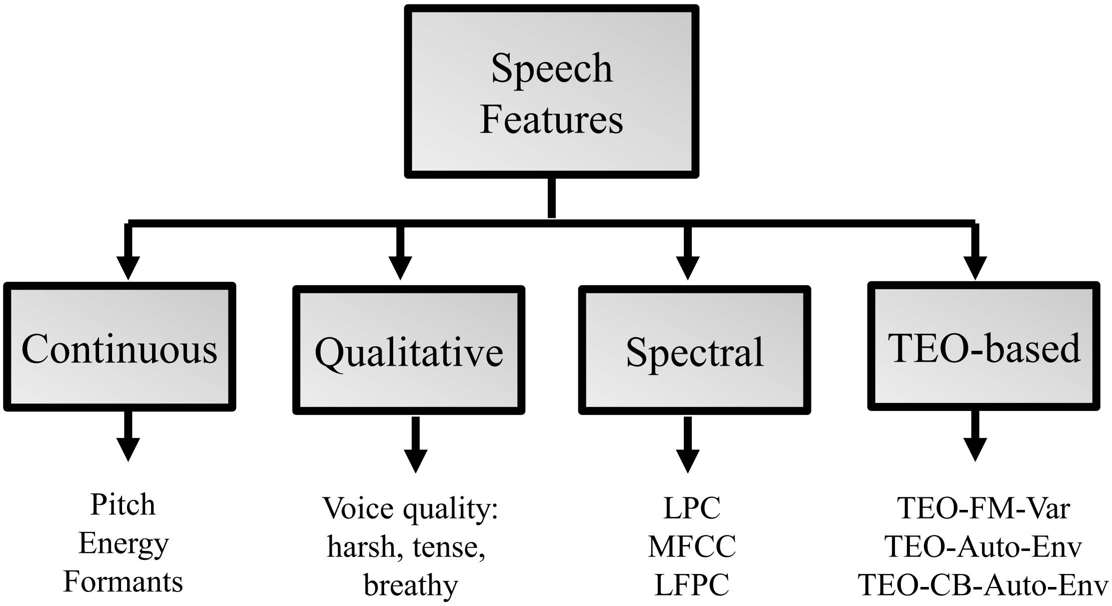


Figure 1 Categories of speech features [8]. Examples of features reported in the literature can be grouped into four categories [8]: continuous features, qualitative features, spectral features, and TEO-based features.

there is ambiguity and subjectivity in the description of voice quality terms such as tense, harsh, and breathy. Various research studies trigger an ongoing debate whether tense voice is associated with anger, joy, and fear; lax voice is associated with sadness and breathy voice is associated with both anger and happiness whereas sadness is associated with a ‘resonant’ voice quality.

In addition to time-dependent features such as pitch and energy, spectral features are often selected as a short-time representation for speech signal. In order to comply with spectral distribution of the auditory system, the estimated spectrum is often passed through band-pass filters or critical bands. The Bark scale, the Mel-frequency scale, the modified Mel-frequency scale, and the ExpoLog scale are the commonly used auditory filter bands. MFCC is a frequently used spectral feature which exploits the human auditory frequency response with the help of Mel-Scale frequency response [12]. It is hard to mention an agreement among the experimental results whether MFCCs achieve poor or robust emotion detection [2,12,20,21]. Alternative to MFCC, the log-frequency power coefficients which include the pitch information are considered for emotion detection as well [6].

The number of harmonics that is produced by the nonlinear air flow in the vocal tract is another useful feature for emotion detection [6,11]. The emotional state of the highly activated modes of anger or stressed speech is caused by the fast air flow which causes vortices located near the false vocal folds providing additional excitation signals other than the pitch. Additional excitation signals in the spectrum are named as harmonics and cross harmonics which form the basis for TEO [2].

It is concluded from the related study that the conventional features have a robust performance in arousal dimension rather than the valence domain. This is because it is confirmed that pitch appears to be an index into arousal [2]. Another well-accepted finding in [2,22] is that mean of the fundamental frequency ($F0$), mean intensity, speech rate, as well as pitch range, blaring timbre, and high-frequency energy are positively correlated with the arousal dimension as well. Shorter pauses and interbreath stretches are indicative of higher activation [9,14]. It is also reported that the unweighted recognition performances achieved for valence mode is highly lower than the arousal. Particularly, TEO features, continuous features such as the fundamental frequency and the pitch features are recommended for classifying high-arousal versus low-arousal emotions [2]. Consequently, it is obvious that there is a strong need for the definition of features which can distinguish different emotions which are arousally similar and valencely different (i.e., angry, fear or sad, bored). Another important issue that needs to be pointed out is the specification of optimal number of features [23]. Existing emotion detection systems mostly employ a variety of features to improve the recognition accuracy without deeply examining the impact of individual features. This is mainly because the conventional features are optimized for speech/speaker recognition rather than the emotional content, hence deal with the structure in spite of perception.

In this article, we propose a new set of acoustic perceptual features for automatic emotion recognition from speech signals. The features are based on the perceptual evaluation of audio quality (PEAQ) standard known from ITU specifications [24,25]. Starting from the outer and middle ear models of the auditory system, we base our features on the masked loudness and perceptual loudness difference concept which defines relatively objective criteria, reference, and emotional difference, for emotion detection. The features include the partial loudness of the emotional difference, emotional difference-to-mask ratio, measures of alterations of temporal envelopes, harmonic structure of the emotional difference, the occurrence probability of emotional blocks, and perceptual bandwidth of emotional audio. It is shown that the novel perceptual features in addition to the introduced reference and emotional difference concepts provide an improved recognition performance in general and particularly for valence.

3. Representation of the emotional content

As it is mentioned in the previous section, among acoustic features energy and loudness play an important role to correlate the speech with the underlying emotion. Most researchers believe that prosody continuous features such as energy and loudness convey much of the emotional content of an utterance [1,12,26].

In the light of this information, we base the principles of our proposed features on the energy and loudness which involves modeling the excitation pattern on the basilar membrane by simulating the acoustic signal transformations in the ear according to the perceptual model of the human auditory system. Unlike the existing features, we emphasize the *perceptual features* which better highlight the perceptual nature of the auditory system [13,24]. This is achieved by detecting the emotional differences which are derived by masking the excitation patterns of emotional signals. In order to take advantage of resemblance, we also make brief comparisons between the aforementioned legacy and the *perceptual features*.

In order to model the emotional content, we introduce nine features that can be categorized as a fusion of the continuous, spectral, and TEO-based features under the perceptual voice quality framework. If we briefly present the features, we propose *perceptual bandwidth* of the emotional signal as a feature that can be linked with the fundamental frequency. The emotional signals alter according to the perceived timbre, dullness, or muffling effects. To measure this effect, we compute a rough estimate of the bandwidth with respect to an adaptively computed noise floor on the high-frequency bands of the spectrum. *Perceptual bandwidth* behaves like sort of an adaptive noise floor threshold finder which makes *perceptual bandwidth* closely related with the overall distribution of the spectrum.

The importance of the harmonics is mentioned in Section 2. In our system, the harmonic structure of the emotional audio is measured with a cepstrum-like analysis which gives a tonal content measure referred as average harmonic structure of magnitude (AHSM). Another important emotional cue is excitation level extracted based on the perceptual masking model. *Average number of emotional blocks* (AEB) as a feature provides a measure for the occurrence of high excitation levels through successive frame groups analyzed in Bark scale. Loudness is also a frequently used asset in speech where our *normalized emotional difference* (NED) feature remarks the “local loudness” level in addition to the “loudness”. Another useful feature is the *normalized spectral envelope* (NSE). The NSE enables us to model the envelope of loudness variations between emotional categories through successive frames.

The main divergence and advantage of our features rely on the perceptual preprocessing steps that are used to model the human auditory system. These perceptual processes cover concepts such as, weighting the spectral components of the audio with the frequency response of the outer and middle ear, using masking models including hearing threshold, and forward and backward masking models. Using these processes, we reformat the energy in a perceptual manner which becomes more compliant to the human auditory system.

The proposed features also have similarities to the MPEG7 descriptors [12] because both intend to represent perceptual audio content thus aim not to rely on speech context. MPEG7 comprises a series of 18 generic LLDs including audio spectrum envelope, harmonic ratio, MFCCs, audio spectrum flatness-centroid-spread, spectral roll of frequency, spectral flux, zero crossing rate, and higher-order statistics derived from the LLDs. The MPEG7 descriptors can be interpreted as a compromised feature set between the conventional features and the perceptual features. However, unlike the proposed features, the MPEG 7 features are derived based on simple hearing threshold of auditory system rather than a perceptual auditory model, hence does not reflect the emotional content efficiently. Detailed formulation of the introduced features is presented in Section 5.

4. Modeling emotional variances

Before giving the formulation of the feature set in this section, we should remark that the subjective nature of the emotion forces researchers to employ a reference criteria which scales the effect of this subjectivity. Conventionally, features such as harmonics-to-noise ratio [27] and preprocessing methods like database normalization involving speaker, corpus, language, or gender are used to cope with the need of defining a reference criteria [14] in emotion detection. It is known that the normalization increases the speaker, corpus, language, or gender dependency while increasing the emotion recognition rate. On the other hand, the measurement of perceptual voice quality differences has been achieved by performing a kind of normalization or noise level estimation to quantify the difference with respect to a reference [24,25].

In a similar manner, we propose the hypothesis that the emotional differences (or variances) are more discriminative than the emotional data itself where the emotional differences are determined through perceptual masking. We can undertake this argument as handling the energy difference between emotional modes is a more robust method in comparison to studying solely on the energy of a single emotional mode. To lay over this theoretical approach on a practical basis, we make use of a *reference* concept to distinguish emotional modes with respect to another emotional category. Our proposed feature vector reflects the variations of every emotional audio sample from the *reference* audio set. The resulting feature vector is called the *perceptual* feature vector since we use ITU perceptual model in feature extraction. The frame-wise computation of the perceptual features is given in Figure 2. Let M denote the dimension of the perceptual feature vector which is set to 9 in this study. In order to efficiently model the characteristic of an emotional mode, the perceptual difference features computed in each audio frame are averaged over Y frames and a single average feature vector is established for the relevant emotional and reference audio signal pair, as it is shown in Figure 2.

In our system, *emotional variances* of both the training and test audio sets are modeled with respect to the same *reference* set. Data belonging to test speaker is not included in the *reference* set, thus there is no need for *a priori* information about all emotional categories of the test speaker. This is an advantage of the proposed method regarding real-time applications. It is also not necessary to use all *reference* records during classification. Although our tests demonstrate that the emotional category of the *reference* set does not play a major role in emotion recognition performance, the *reference* audio

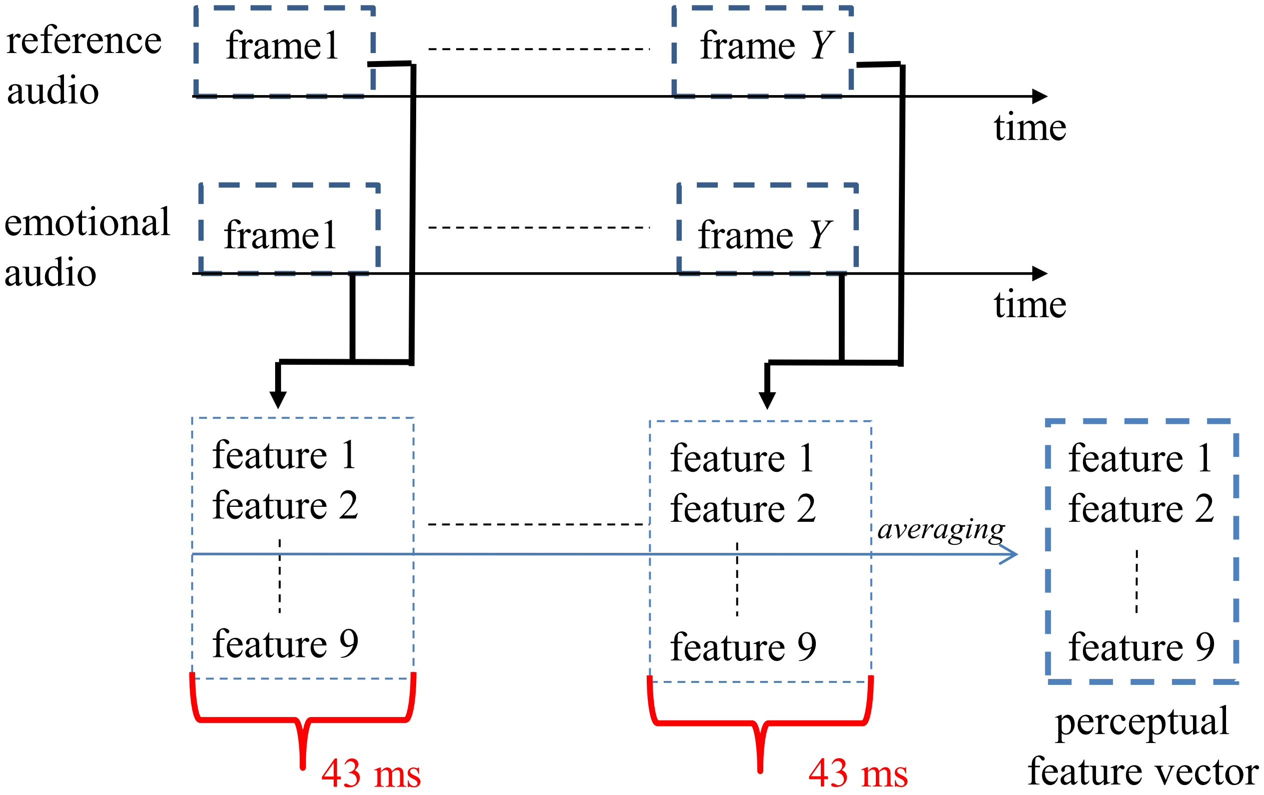


Figure 2 Frame-based perceptual feature extraction. In order to efficiently model the characteristic of an emotional mode, the perceptual features computed in each frame are averaged over *Y* overlapping frames. Hence, a single *M*-dimensional average feature vector is extracted from each emotional and reference audio pair and is used as a training or test vector.

set is chosen in such a way that to highlight between class variations. Note that size of the reference set should be appropriate as in the case of all training practices. The role of the reference set is depicted in Figure 3. Let *N* be the size of the reference set. An emotional utterance, i.e., an utterance with mode angry shown in Figure 3, is modeled by *N* feature vectors each including *M* features.

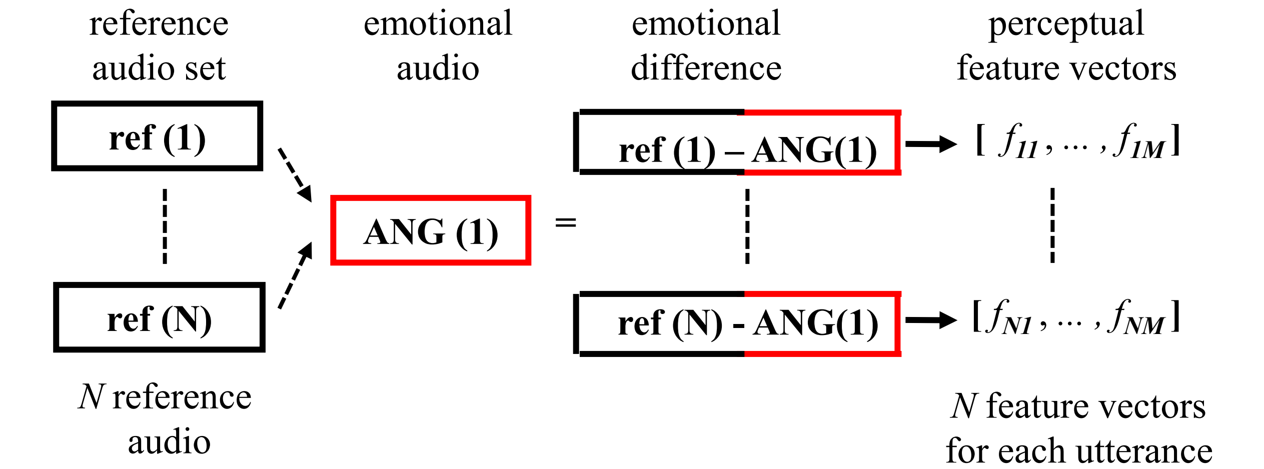


Figure 3 Coupled reference set and the perceptual features. An utterance is represented by *N* perceptual feature vectors each extracted with respect to a reference audio sample. Hence, the perceptual feature vectors reflect the variation of every emotional audio sample from the reference set.

The perceptual vectors representing each utterance are exposed to the SVM/GMM classifier and each test utterance is classified on the assessment of the posterior probabilities from the emotional classes. The architecture of the proposed system is shown in Figure 4. In our previous study [13], we applied majority voting to improve the recognition rate by the way of maximizing the probability on the instance base. The drawback with this procedure is the partial utilization of the posterior probabilities as the category probabilities are close to each other.

In this study, we propose a decision rule which can be interpreted as an S-MV rule. The new approach enriches the support factors for decision boundaries thus providing a more robust and immune structure involving the selection of the reference set. Equations (1) to (4) define the new decision rule stemming from the statistics of the posterior probabilities generated by the SVM/GMM classifier. Let j be the index for emotion categories where j = 1,2 for pairwise classification that we employed. w_j stands for the class label j, x^i is the ith perceptual feature vector where $p(w_j|x^i)$ is the conditional probability of feature vector x^i being classified with class label w_j.

We decide on a final score in S-MV for emotional labels. We try to force the achieved conditional probabilities $p(w_j|x^i)$ to a higher level with a decision rule. $P^i_{j\mathrm{final}}$ in Equation (1) is the trivariate decision metric, which provides the final score of decision

$$P^i_{j\mathrm{Final}} = P^i_{jPr} + P^i_{j\mathrm{Var}} + P^i_{jV} \tag{1}$$

The final score is computed for each category j based on feature vector x^i of each utterance. The category with the highest score $P^i_{j\mathrm{Final}}$ is assigned to the vector x^i. In the breakdown of the final score, P^i_{jPr} shown in Equation (2) governs the mean of the conditional probabilities where N is the size of the reference set. This subcriterion enables the emotion labeling with the higher posterior probability

$$P^i_{jPr} = \frac{1}{N}\sum_{i=1}^{N} p(w_j|x^i), \tag{2}$$

where j = 1, 2, and $\sum_j P^i_{jPr} = 1$.

We also favor perceptual vectors with the lower variance value, $P^i_{j\mathrm{Var}}$ as given in Equation (3), where ξ_j is the variance of $p(w_j|x^i)$ and T is the sum of all the variance values through j categories. Thereby, we reward the category having lower variance among the conditional probabilities generated by SVM/GMM

$$P^i_{j\mathrm{Var}} = 1 - \frac{\xi_j}{T}, \tag{3}$$

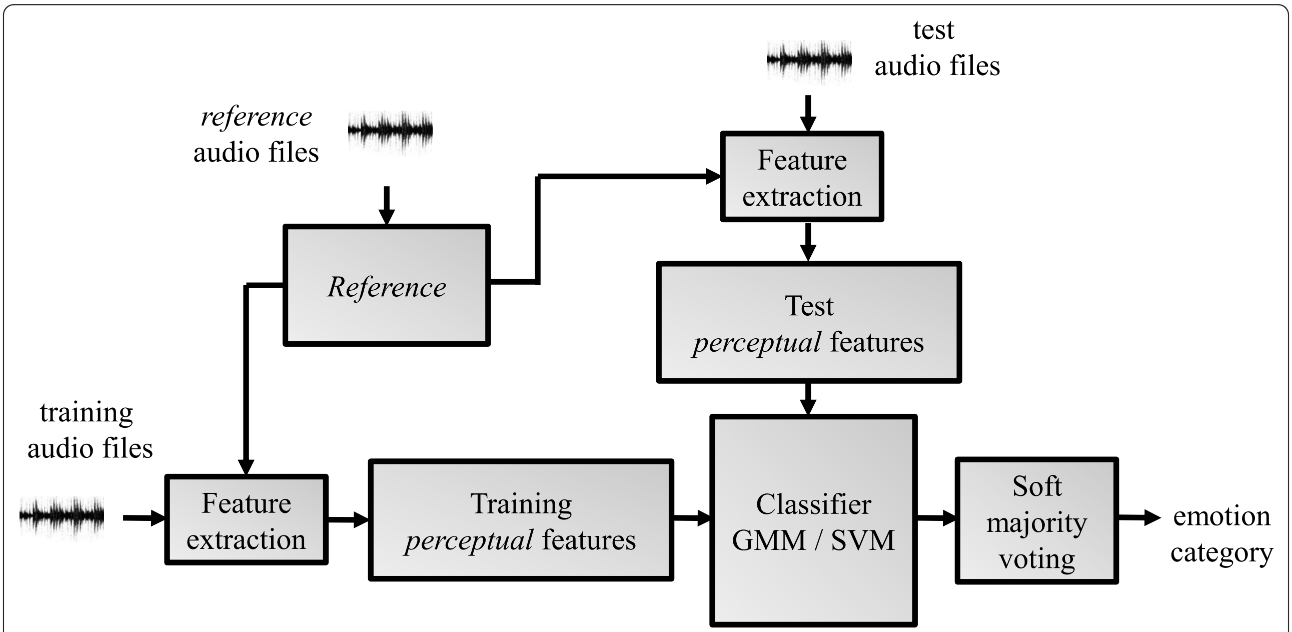


Figure 4 Architecture of the perceptual feature extraction and the classification system. The reference set is shared by both the training and the test datasets. For each training/test audio utterance, *N* perceptual vectors are extracted and fed into a classifier. Classifier design has been performed based on the perceptual features extracted from the training data. Similarly, the test features are derived from the test data. S-MV has been used for the category labeling.

where

$j = 1, 2, \quad \xi_j = \frac{1}{N} \sum_{i=1}^{N} \left(p(w_j|x^i) - P^i_{jPr} \right)^2 T = \sum_{j=1}^{2} \xi_j,$

and $\sum_j P^i_{j\mathrm{Var}} = 1.$

However, our study revealed that additional mechanisms are necessary to avoid the confusion on the category borders, particularly for valence. The primary need comes from the limited size of the reference set N referring to the sparse data. The voting score P_{jV} in Equation (4) applies a voting system which makes use of the relative probability values of the categories by a quantification linear function $g(p(w_j|\ x^i))$.

$$P^i_{jV} = \frac{1}{N} \sum_{i=1}^{N} g\left(p(w_j|x^i)\right), \tag{4}$$

where $g\ \left(p(w_j|x^i)\right) = \begin{cases} 1 & \arg\left\{\max\limits_j \left(p(w_j|x^i)\right)\right\} \\ 0 & \text{otherwise} \end{cases}$ and $\sum_j P^i_{jV} = 1$

The size of the reference set plays an important role for the S-MV. As the size of the reference set N increases, the amount of statistical data explored is increased as well thus more precise performance can be reported. On the other hand, higher the N, higher the computational complexity. Hence, for each corpus, we have specified the N heuristically considering this tradeoff.

We call this overall approach *S-MV* that takes greater amount of statistical data into consideration with regard to applying solely *majority voting* as in our previous study [13].

5. Formulation of the perceptual features

In this section, we first give the notation and summarize the basic preprocessing steps of perceptual masking. Later, we present the mathematical formulation to describe how we derived each feature.

5.1. Preprocessing

In the preprocessing step, time signals sampled at 16 kHz are divided into frames of 43 ms with 50% overlap where the emotional audio signal behaves in a stationary manner by which we can better model its statistics. Let $s_n[k_t, n]$ denote the temporal domain signal where n is the index of time-frames and k_t is the time sample index. In order to apply short-time Fourier Transform (STFT), the windowed audio frame can be represented as

$$s_w[k_t, n] = h_w[k_t] s_n[k_t, n], \tag{5}$$

by using the Hann window

$$h_w[k_t] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - \cos\left(\frac{2\pi k_t}{N_{FT} - 1}\right)\right] \ , \tag{6}$$

where N_{FT} is the is the size of the DFT which is equal to 2048 in this study.

Successive frames of the time-domain signal are transformed to a basilar membrane representation based on the PEAQ psycho-acoustic model [24,25]. Hence, first each windowed frame is transformed to the frequency domain by taking STFT. Let the transformed and windowed audio frame be expressed as

$$F[k_f, n] \ = \ \frac{1}{N_{FT}} \sum_{k_t = 0}^{N_{FT}-1} s_w\left[k_t, \ n\right] \mathrm{e}^{-j \frac{2\pi}{N_{FT}} k_f k_t}, \tag{7}$$

where k_f is the frequency bin index. In order to extract the perceptual components of the audio spectrum, a mapping reflecting the outer and middle ear frequency responses is applied on the spectral components, yielding the "*Outer ear weighted DFT outputs*" given as

$$F_{\mathrm{e}}[k_{\mathrm{f}}, n] = \left|F[k_{\mathrm{f}}, n]\right| \cdot 10^{\frac{W[k_{\mathrm{f}}]}{20}}. \tag{8}$$

The weighting function $W[k_{\mathrm{f}}]$ shown in Equation (8) represents the effect of the ear canal and the middle ear frequency response [24,25]. The outer middle ear frequency response is formulated as

$$W[k_{\mathrm{f}}] = -0.6 \cdot 3.64 \cdot {k_{\mathrm{f}}}^{-0.8} + 6.5 \cdot \mathrm{e}^{-0.6\cdot(k_{\mathrm{f}}-3.3)^2} - 10^{-3} \cdot (k_{\mathrm{f}})^{3.6} \tag{9}$$

where k_{f} denotes the frequency bin index.

Note that these weights enable us to filter the spectral components according to the human auditory system, because both the outer and middle ears act as band pass filters. Hence, the outer and middle ear transfer functions limit the ability to detect low-amplitude audio signals and affect the absolute threshold of hearing [24,25]. According to the frequency response of the outer and middle ears, the absolute threshold of hearing tend to be lowest in the 2-3 kHz band and increases with increased or decreased frequency. The frequency components lower than 1 kHz are drastically attenuated and the components between 3 and 4 kHz become stronger and perceived better. The frequency borders of the band pass filter range from 80 to 18000 Hz.

Unlike the conventional audio feature extraction modules that mostly operates in Mel scale, in which speech contents are efficiently modeled rather than emotion, we propose working on perceptual spectrums derived in Bark scale. Hence, a mapping from the frequency domain to Bark scale is performed. Such an approximation leads to the notion of critical bands or perceptual scales in other words. Hence, the frequency bins of the attenuated spectral energy values are grouped into z = 109 critical bands as in the basic version of the PEAQ model. The attenuated spectral energy values are mapped from frequency domain and grouped into a pitch (Bark) scale by the following approximation as

$$z(\text{Bark}) \approx 7 \cdot \operatorname{arsinh}\left(\frac{f}{650}\right). \tag{10}$$

It can be seen from Equation (10) that the Bark scale frequency bands are almost linear below 1 kHz while they grow exponentially above 1 kHz that yields a perceptual filter bank.

Let $F_{\mathrm{p}}[k_{\mathrm{f}},\ n]$ be the energy representation of the "Outer ear weighted FFT outputs" and $P_{\mathrm{e}}[k,\ n]$ is the Bark representation of $F_{\mathrm{p}}[k_{\mathrm{f}},\ n]$ = $|F_{\mathrm{e}}[k_{\mathrm{f}},\ n]|^2$. Note that the frequency index k_{f} in Hz is replaced by k after mapping to Bark scale. The energy components which are transformed to Bark domain are convolved with a spreading function $S_{\mathrm{dB}}(.)$ to simulate the dispersion of energy along the basilar membrane and to model the spectral masking effects in the Bark domain. The pitch patterns, $P_{\mathrm{e}}[k,\ n]$, are smeared out over frequency using the level dependent spreading function. Conventionally, $S_{\mathrm{dB}}(i,\ k,\ n,\ P_{\mathrm{e}})$ the spreading function of band i for an energy component at band k is defined as a two-sided exponential

$$S_{\mathrm{dB}}(i,k,n,P_{\mathrm{e}}) = \begin{cases} 27(i-k)\Delta z & ;k < i \\ \left[-24 - \dfrac{230}{f_c} + 2\log_{10}\left(P_{\mathrm{e}}(k,n)\right)\right](i-k)\Delta z & ;k > i \end{cases}, \tag{11}$$

where Δz = i-k = 1/4 for the basic version of PEAQ. Smearing the spectral energy over frequency gives the frequency domain spreading function, $E_{\mathrm{s}}[k,\ n]$ which is called as the "unsmeared excitation pattern" [24,25],

$$E_{\mathrm{s}}[k,n] = \frac{\left(\sum_{k=0}^{N_c-1} P_{\mathrm{e}}[k,n] S_{\mathrm{dB}}(i,k,n,P_{\mathrm{e}}[k,n])^{0.4}\right)^{\frac{1}{0.4}}}{B_{\mathrm{s}}[i,k,n]}, \tag{12}$$

where $B_{\mathrm{s}}[i,\ k,\ n]$ is a normalizing factor which is calculated for a reference level of 0 dB and can be pre-computed since it does not depend on the data.

The feature extraction process is then followed by a time domain spreading that accounts for forward masking effects. In spite of the conventional time masking functions commonly used in audio compression, we prefer using the one introduced in PEAQ that enables us tracking emotional variances of successive frames. Hence, to model forward masking, the energy levels in each critical band are smeared out over time according to Equation (13) as

$$\bar{E}[k,n] = a \cdot \bar{E}[k,n-1] + (1-a) \cdot E_{\mathrm{s}}[k,n]^{0.3}, \tag{13}$$

where a is a time constant depending on the center frequency of each critical band. The *excitation pattern*, $E[k,\ n]$, shown in Equation (13) is calculated as

$$E[k,n] = \max\left(\bar{E}(k,n), E_{\mathrm{s}}(k,n)\right), \tag{14}$$

where n is the actual frame number, k is the band index and $\bar{E}[k,0] = 0$.

Briefly, we observed the perceptual effect of our method in means of both the proposed features and the perceptual intermediate steps which are applied prior to feature extraction. In Figure 5, the impact of the preprocessing steps is provided comparatively for two audio records taken from VAM. Note that perceptual masking in Bark scale highlights the emotional differences.

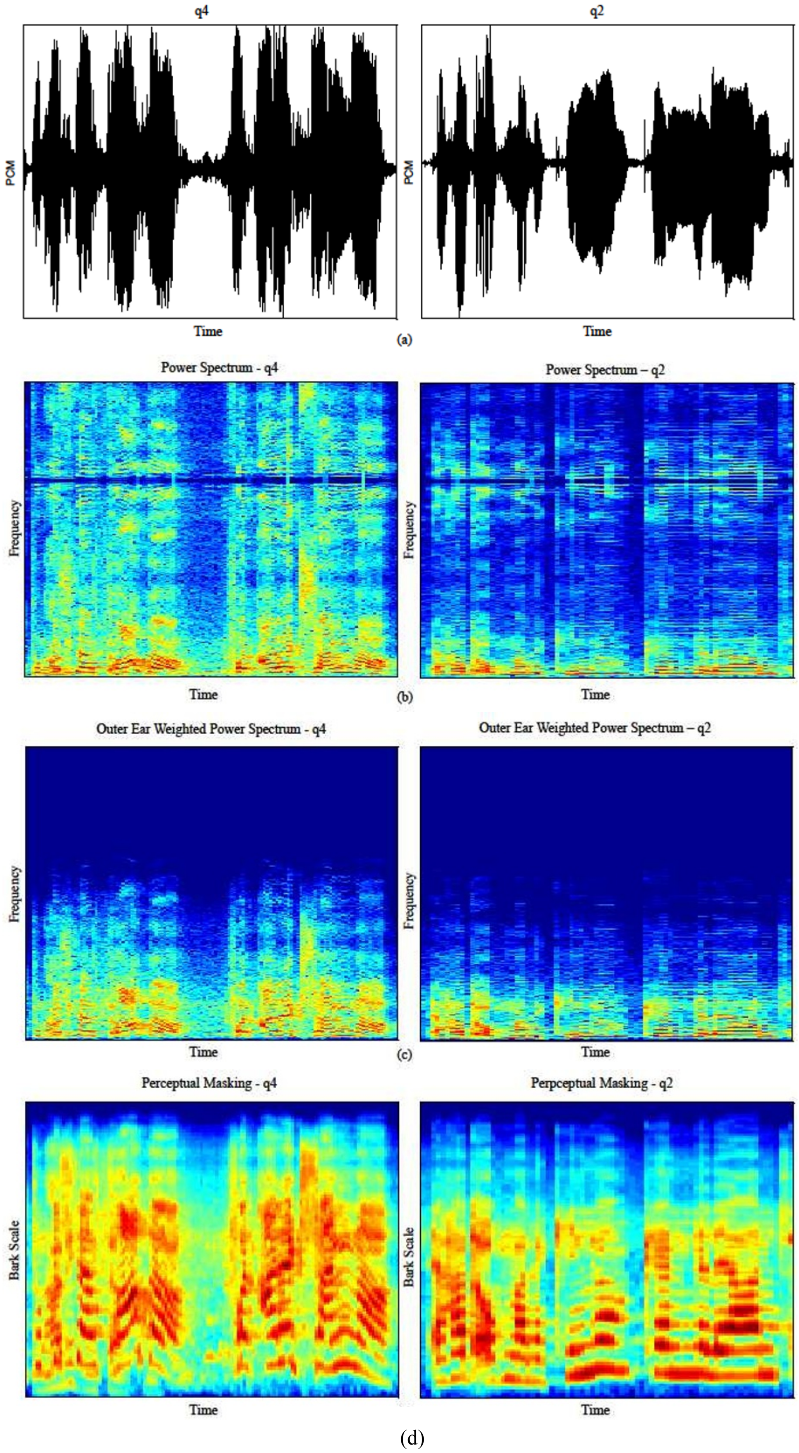
5.2. Perceptual features

The preprocessing stages detailed in the previous section are employed in the system in order to model physiological and perceptual effects of the human ear. The preprocessing is followed by feature extraction process. The proposed perceptual feature set consists of seven low level and two statistical descriptors. These features are formulated in the following sections.

5.2.1. Average harmonic structure magnitude of the emotional difference

Our motivation of using AHSM of the *emotional difference* as a representative feature is to highlight the harmonic structure of emotional speech that is much more similar to a periodical signal with stable harmonics with respect to unemotional speech. Depending on the fact that emotion does not change as fast as the phonemes, we prefer to use the correlation of the logarithmic power spectrum rather than the spectrum of signal itself. So, AHSM emphasizes the harmonic pattern and reflects the variations in the fundamental frequency. AHSM is a useful feature to discriminate the speech on the valence dimension where a sample for q2-q3 pairwise discrimination is shown in Figure 6.

The harmonic structure of the signal is evaluated in linear frequency spectrum rather than Bark, since nonlinear frequency transformation would smear the harmonic structure. Extraction of the feature AHSM can briefly be summarized as follows. First, we compute the emotional differences for each critical band. Then, the autocorrelation function of emotional differences through the critical bands is obtained. Fundamental frequency is estimated from the log-spectrum of the autocorrelation function. Average value of the fundamental



(d)

Figure 5 The perceptual preprocessing steps in various emotional modes. Prior to feature extraction, we perform preprocessing on utterances to highlight the perceptual content of the audio. To make this procedure apparent, q4 (high arousal) (first column) and q2 (low arousal) (second column) modes are considered from VAM. The first two rows illustrate time domain and power spectrum of the files taken from, respectively, q4 and q2 categories. The spectral properties are shaped by an auditory filter bank, resembling the hearing threshold of the outer ear (third row). The disparity stemming from the emotional characteristics improve after perceptual masking performed in Bark domain (fourth row).

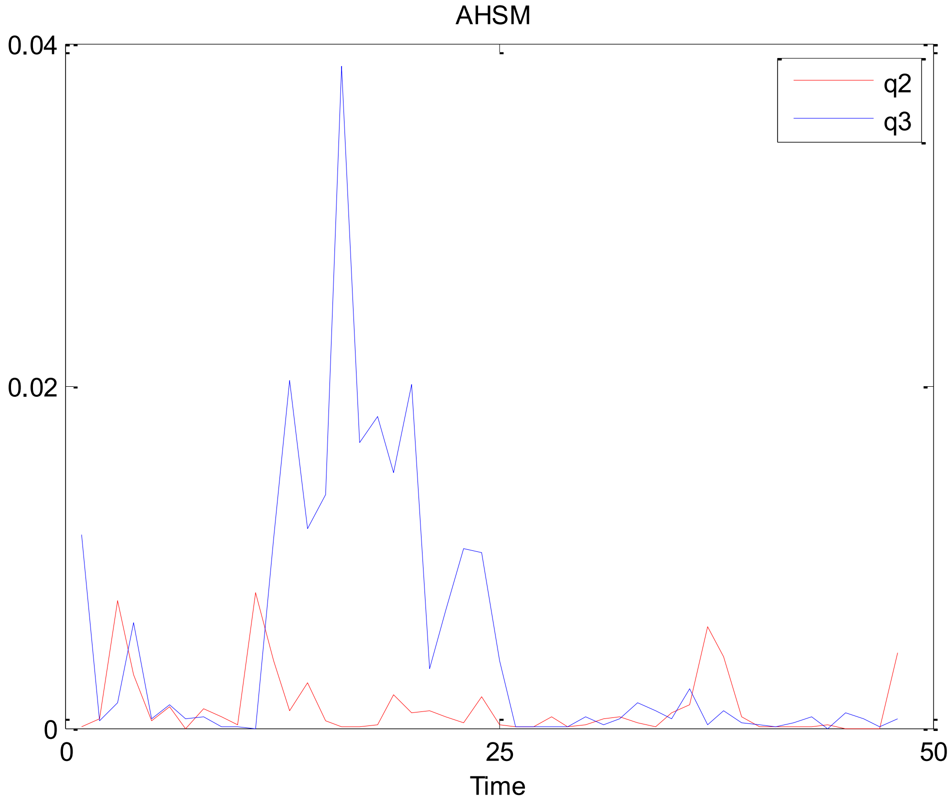


Figure 6 Distribution of the feature AHSM versus time for two audio utterances sampled, respectively, from q2 and q3 modes of VAM [16]. q2 and q3 locate at the positive and the negative sides of the valence dimension, respectively. It can be observable that the feature AHSM provides a clear distinction between q2-q3 on the valence dimension. q3 is more likely to have tonation fluctuations in negative valence mode (i.e., sadness) which avoids q3 mode to have a harmonic structure as smooth as the q2 signal (i.e., neutral). The periodicity in q2 plainly appears from AHSM values in contrary to the fluctuating harmonics of q3. If the speech utterance has a periodicity, it can be evaluated based on the AHSM.

frequencies estimated for successive Y audio frames is reported as AHSM. Conventionally, fundamental frequency is estimated from the log-spectrum of autocorrelation function of audio signal [6]. Unlike these methods, we use the correlation of emotional differences through critical bands instead of time domain audio signal itself.

To formulate the AHSM, let the emotional difference $P_{\mathrm{EDiff}}[k_{\mathrm{f}}, n]$ of frame n in spectral index k_{f} refer to the variations from the reference set within that band. $P_{\mathrm{EDiff}}[k_{\mathrm{f}}, n]$ in Equation (15) is calculated in the frequency domain as the log spectra of the ratio of magnitudes of the emotional and the reference audio signals spectral energy $F_{\mathrm{eE}}[k_{\mathrm{f}}, n]$ and $F_{\mathrm{eR}}[k_{\mathrm{f}}, n]$, respectively. Note that the spectral energy obtained after outer and middle ear filtering of the STFT spectrum is calculated by Equation (8)

$$P_{\mathrm{EDiff}}[k_{\mathrm{f}}, n] = \log\left(\left|F_{\mathrm{eE}}[k_{\mathrm{f}}, n]\right|^2\right) - \log\left(\left|F_{\mathrm{eR}}\ [k_{\mathrm{f}}, n]\right|^2\right) = \log\frac{\left|F_{\mathrm{eE}}[k_{\mathrm{f}}, n]\right|^2}{\left|F_{\mathrm{eR}}[k_{\mathrm{f}}, n]\right|^2} = P_{k_{\mathrm{f}}}. \quad (15)$$

Let $P_{k_{\mathrm{f}}}$ denote the row vector obtained by grouping the energies of the emotional differences at critical bands k_{f} through successive 256 frames. The normalized autocorrelation function $C[l]$ is calculated in a defined neighborhood of $l = 256$. Hence, the normalized autocorrelation function $C[l]$ of gathering emotional differences in M groups within a defined neighborhood l is calculated as

$$C[l] = \frac{P_{k_{\mathrm{f}}}.P_{k_{\mathrm{f}}+l}}{\sqrt{\left|P_{k_{\mathrm{f}}}\right|^2\left|P_{k_{\mathrm{f}}+l}\right|^2}}. \quad (16)$$

The power spectrum $S[k_{\mathrm{f}}]$ of the normalized autocorrelation function is calculated by

$$S[k_{\mathrm{f}}] = \left|\frac{1}{256}\sum_{i=1}^{256} C[l]e^{j2\pi k_{\mathrm{f}} l/256}\right|^2, \quad (17)$$

and its maximum peak specifies the harmonic magnitude $E_{H\mathrm{max}}[n]$. AHSM is the average of the magnitudes estimated through Y successive audio frames and calculated as

$$\mathrm{AHSM} = \frac{1000}{Y}\sum_{n=0}^{Y-1} E_{H\max}[n] \qquad (18)$$

5.2.2. Average number of emotional blocks

The excitation patterns of different emotional audio signals are processed and stored in the brain. The brain keeps the brief initial audio information in a short-term memory. Subjective evaluation of the emotional signals depends on this short-term memory [25]. Hence, the feature AEB provides a measure for the occurrence of high excitation levels through successive frame groups analyzed in Bark scale. To calculate the expected number of emotional blocks within a time interval, a probabilistic approach that estimates the number of excitation patterns remaining above a loudness threshold is applied [24].

AEB provides a measure for the occurrence of high excitation levels through successive Y frames analyzed in Bark scale. Specification of Y is directly related to the granularity of the system and it is set to $Y = 70$ in this study. In order to calculate the expected number of emotional blocks within a time interval, we have applied a probabilistic approach that estimates the number of excitation patterns remaining above a loudness threshold.

Let $e[k, n]$ denote the difference between the excitation levels of *reference* and emotional audio computed in Bark scale k for audio frame n in dB as

$$e[k,n] = 10\log_{10}\left(\frac{E_{sE}[k,n]}{E_{sR}[k,n]}\right). \qquad (19)$$

Our aim is specifying frames in which the excitation level difference above a threshold [24,25]. Probability of an excitation pattern remaining above a loudness threshold can be modeled by [24]

$$p[k,\ n] \ = \ 1 \ - \ \left(\frac{1}{2}\right)^{\left(\frac{e[k,n]}{s[k,n]}\right)^{b}}, \qquad (20)$$

where b is a constant equal to 6 and where $s[k, n]$ is a normalizing coefficient. Hence, assuming that the observed frames are uncorrelated, the total probability of declaring the frame n as emotional can be calculated by

$$P[n] \ = \ 1 - \prod_{\forall k} \left(1 - p[k,n]\right). \qquad (21)$$

Basically, the feature AEB is computed as the average number of blocks declared as emotional within 1 s. It can be shown that $P[n]$ becomes greater than 0.5 for these frames. Since both probability of detection and number of steps remaining above the loudness threshold are dependent on the excitation patterns, we can expect the excitation pattern of the audio in mode happy to have higher peaks with respect to the mode bored which are, respectively, located on the positive and negative scales of arousal. The discrimination capability of the feature AEB is promising as it can be seen in Figure 7 for the pairwise training set of bored and happy modes.

5.2.3. Perceptual bandwidth

The perceptual bandwidth of emotional audio varies according to the perceived timbre, dullness, or muffling effects. To measure this effect, the maximum of the frequency spectrum in upper frequency range is obtained. This is used as an estimate of the noise floor. Then, beginning from higher frequencies and scanning the highest frequency component which exceeds the noise floor by at least 10 dB toward lower frequencies is defined as the estimated perceptual bandwidth. This feature aims to classify emotional states based on the variations in signal bandwidth. Hence, a rough estimate of the observed emotional signal bandwidth is computed for each audio frame with 43 ms. To do this, first the maximum $W_1[n]$ of the spectrum of the emotional audio signal obtained within a frequency band from 14.4 to 16 kHz is specified as the noise floor by using

$$W_1[n] = \max_{k_f}\left\{F[k_f,n]\right\}, \quad 14.4\,\mathrm{kHz} < k_f < 16\,\mathrm{kHz}. \qquad (22)$$

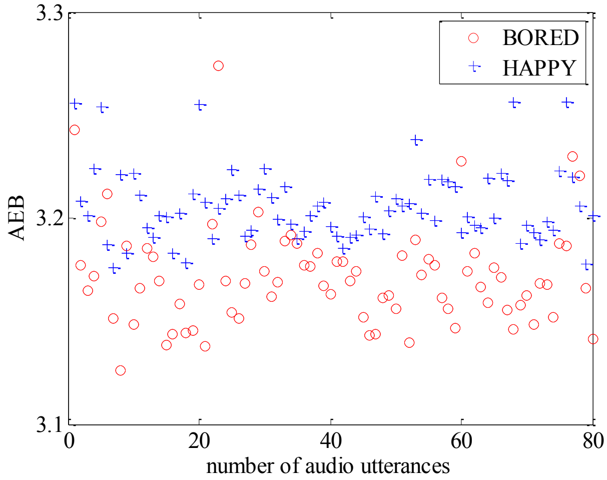


Figure 7 A pairwise distinction for bored and happy modes from EMO-DB [15]**can be modeled by the feature AEB**. The features have a heterogeneous effect on emotional modes in pairwise classification. A sample distribution of the perceptual feature is given where AEB plays an important role in bored-happy pairwise classification.

The first frequency component where the spectral energy exceeds the noise floor at least by 10 dB in the reference audio signal is reported as the bandwidth of the emotional audio for the nth frame and it is denoted by $W_2[n]$ and calculated as

$$W_2[n] = \arg_{k_f}\left\{F[k_f, n] > 10\log\left(F\left[W_1[n],n\right]\right)\right\}, \quad 3\,\text{kHz} < k_f < 14.4\,\text{kHz}. \tag{23}$$

Furthermore, searching downward from $W_2[n]$, the first value which exceeds $W_1[n]$ by 5 dB in the emotional audio signal is recorded as $W_3[n]$

$$W_3[n] = \arg_{k_f}\left\{F[k_f, n] > 5\log\left(F\left[W_2[n],n\right]\right)\right\} \\ \left\{F[k_f, n] > 5\log\left(F\left[W_2[n],n\right]\right)\right\}, \quad k_f < W_2[n]. \tag{24}$$

The perceptual bandwidth of the emotional audio is extracted by calculating the mean value over Y successive frames as

$$W_E = \frac{1}{N}\sum_{n=0}^{Y-1} W_2[n] \tag{25}$$

Discrimination capability of the feature W_E can be seen from Figure 8 that plots the distribution of bandwidth estimates through the samples taken from q4 and q1 modes. Perceptual bandwidth of the reference audio is extracted by calculating the mean value over Y successive frames as

$$W_{E1} = \frac{1}{N}\sum_{n=0}^{Y-1} W_3[n] \tag{26}$$

5.2.4. Normalized spectral envelope

The term spectral envelope refers to the normalized amplitude variations of loudness that arise from the emotional differences of successive frames. NSE NSE[k, n] formulated in Equation (27)

$$\mathrm{NSE}[k,n] = \frac{\bar{E}_{\mathrm{der}}[k,n]}{1 + (\bar{E}[k,n]/0.3)}, \tag{27}$$

and NSE difference NSEDiff[k, n] given as

$$\mathrm{NSEDiff}[k,n] = w \cdot \frac{\left|\mathrm{NSE}_E[k,n] - \mathrm{NSE}_{R[}k,n]\right|}{\beta + \mathrm{NSE}_E[k,n]} \tag{28}$$

to quantify local variations in energy through time. $\bar{E}_{\mathrm{der}}[k,n]$, shown in Equation (27), is calculated by

$$\bar{E}_{\mathrm{der}}[k,n] = a \cdot \bar{E}_{\mathrm{der}}[k,n-1] + (1-a)\cdot \left|E_S[k,n]^{0.3} - E_S[k,n-1]^{0.3}\right|, \tag{29}$$

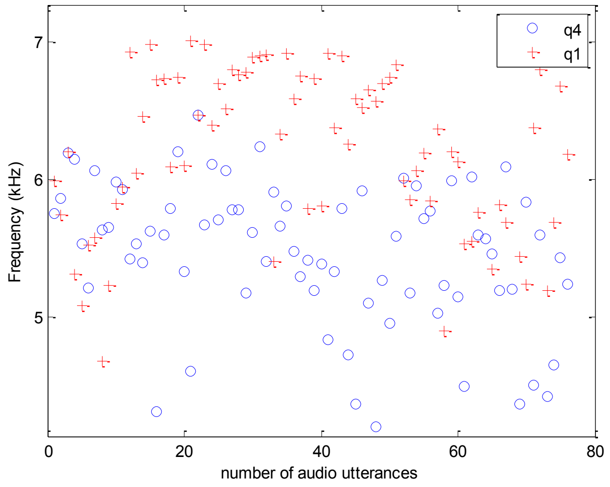


Figure 8 Effectiveness of the feature perceptual bandwidth on the pairwise distinction of modes q1 and q4 from VAM [16]. The perceptual bandwidth is closely related with the overall distribution of the spectrum. If spectral components of the emotional signal are dominant at high-frequency bands, the bandwidth tends to be at the lower frequencies because the noise floor is high as well. Otherwise, if the emotional signal has dominant lower-frequency components, then the noise floor tends to be lower and the bandwidth parameter probably takes a relatively greater value. Perceptual bandwidth behaves like sort of an adaptive noise floor threshold finder. Also shows that q4 and q1 are positively arousal modes. q4 which is in the negative valence domain owns a greater noise floor that yields a lower perceptual bandwidth with regard to q1 signal. As the perceptual bandwidth range of the q1 signal is mostly on the 6-7 kHz band where q4 signal is scattered on the 5-6 kHz band.

where $E_S[k, n]$ is the unsmeared excitation pattern formulated in Equation (12).

As it can be seen, $\bar{E}_{\mathrm{der}}[k,n]$ models the envelope changes through successive frames. The parameter a ($0 < a < 1$) shown in Equation (29) reflects the impact of the past frames to current nth frame gradually. The frame which has maximum effect on $\bar{E}_{\mathrm{der}}[k,n]$ other than nth frame is the previous (n - 1)th frame. The scalar a behaves like an attenuation parameter which both reflects the impact of the past and reduces the effect of the past gradually. This concept precisely models the time spreading energy effect which is based on the perceptual effect of the sequential speech tones. A sample case is given in Figure 9 for angry and bored modes.

The term $\bar{E}_{\mathrm{der}}$, $\bar{E}[k,n]$ shown in Equation (27) denotes the average loudness that effects as an adaptive normalization term on the loudness and is calculated according to

$$\bar{E}[k,n] = a \cdot \bar{E}[k,n-1] + (1-a)\cdot E_S[k,n]^{0.3}. \tag{30}$$

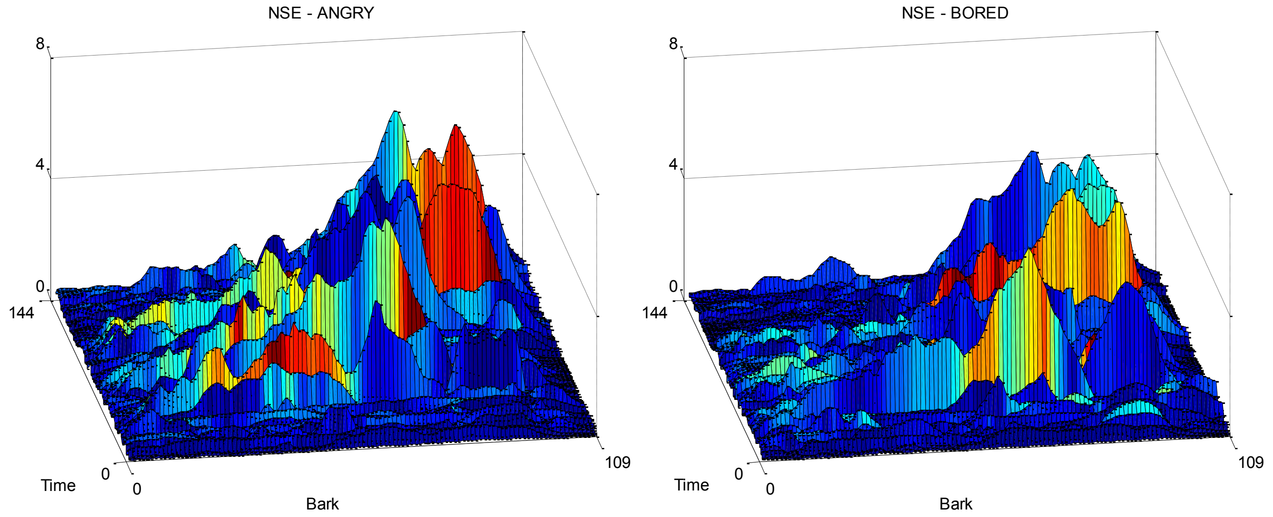


Figure 9 Distribution of the feature NSE [*k*, *n*] in time-frequency domain for angry and bored modes from EMO-DB [15]. The normalized amplitude variations of loudness in angry mode are expected to be higher because of sudden rise and falls in the utterance. The distribution clearly shows that the NSE[*k*, *n*] between successive time frames through the Bark scales are higher for the mode angry.

5.2.5. Normalized emotional difference

The perceived loudness of an emotional audio signal depends on its duration and its temporal and spectral structure. The local loudness of an emotional signal is the perceived loudness after it has been reduced by a masker [25]. The masker induces the loudness to be perceived at different frequency bands. The masker is effective in the low frequencies, thus the locality is established by adaptively masking the low-frequency components. This masking describes the effect by which an audible signal becomes inaudible when a louder signal masks it. We refer the *reference* audio signal as masker and compute a local loudness other than conventional loudness values. In conclusion, we evaluate a localized loudness with respect to a *reference* set.

The NED is formulated as the ratio of the emotional difference $P_{\mathrm{EDiff}}[k_{\mathrm{f}}, n]$ given by Equation (15) to the masking threshold $M[k, n]$. We use the total NED that is calculated as the average (expressed in dB) of the NED values computed at the bark scales,

$$\mathrm{NED}_{\mathrm{tot}} = 10\log_{10}\frac{1}{Y}\sum_{n=1}^{Y}\left(\frac{1}{Z}\sum_{k=0}^{z-1}\frac{P_{\mathrm{EDiff}}[k,n]}{M[k,n]}\right), \tag{31}$$

where $Z = 109$, k denotes the number of critical bands and n refers to the frame number.

The masking threshold $M[k, n]$ formulated below

$$M[k,\ n] = \frac{E[k,\ n]}{10^{\frac{m[k]}{10}}} \tag{32}$$

is calculated by weighting the excitation patterns $E[k, n]$ with the masking offset $m[k]$ as given in

$$m[k] = \begin{cases} 3.0 & \text{for } k \cdot \Delta z \leq 12 \\ 0.25 \cdot k \cdot \Delta z & \text{for } k \cdot \Delta z > 12 \end{cases}. \tag{33}$$

The masking offset is plotted in Figure 10. Since masking offset is placed at the denominator of both the masking threshold, $M[k, n]$, and the NED, it effects as a high pass filter. Hence, the expectation from the feature NED is to emphasize the distinction between emotional categories at higher frequencies as shown in Figure 11.

5.2.6. Emotional loudness

We propose using the overall loudness of the emotional differences as a representative feature of the emotional modes. The specific loudness pattern for a signal can be formulated as

$$L[k,n] = \mathrm{const}\cdot\left(\frac{1}{s[k]}\cdot\frac{E_{\mathrm{IN}}[k]}{10^4}\right)^{0.23}\cdot\left[\left(1-s[k]+\frac{s[k]\cdot E[k,n]}{E_{\mathrm{IN}}[k]}\right)^{0.23}-1\right], \tag{34}$$

where E_{IN} is the internal noise of the ear. The threshold index $s[k]$ is calculated according to

$$s[k] = 10^{\frac{1}{10}\left(-2-2.05\cdot atn\left(\frac{f}{4000}\right)-0.75\cdot atn\left(\left(\frac{f}{1600}\right)^2\right)\right)}. \tag{35}$$

The overall loudness of the signal L_{total} is calculated as the sum across all filter channels of all specific loudness values above zero, as

$$L_{\mathrm{total}}[n] = \frac{24}{Z}\cdot\sum_{k=0}^{Z-1}\max\left(L[k,n],0\right). \tag{36}$$

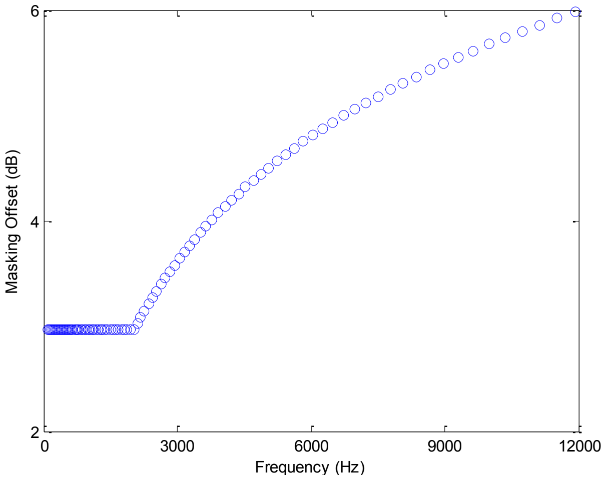


Figure 10 The masking offset function as a function of frequency. The masking offset behaves linearly until 2 kHz while it augments the spectral components above 2 kHz. The markers in the figure indicate the Bark center frequencies [24].

An example of the effect captured with this feature is shown in Figure 12.

6. Test results

The popular emotion databases cover a vast diversity beginning from naturalistic and spontaneous datasets in comparison to acted datasets which contain pre-defined phrases spoken by actors and actresses in a studio environment. It is obvious that the naturalistic datasets address the realistic problems of emotion recognition in everyday life. Among a number of databases, we have used well-known EMO-DB [15] and VAM [16] as, respectively, acted and spontaneous database. EMO-DB corpus covers pre-defined sentences spoken by ten actors in seven emotions; *anger, boredom, disgust, fear, joy, neutral, and sadness*. VAM consists of audio recordings from a German TV show including spontaneous and emotionally colored phrases from 47 guests. Being different from EMO-DB, VAM data are labeled by 17 human labelers on a 5-point scale for three dimensions (valence, activation, and dominance). The applied mapping method given in Figure 13 is a biased assumption other than being straight forward. The assumption is required from the need of standardizing diverse emotions of different databases.

The weighted (WA) and unweighted (UA) average of class-based recall rates are assessed as evaluation measures on the pairwise multiclass discrimination. EMO-DB and VAM covers a total of seven and four emotional categories, respectively. *Number of audio* utterances used in training and test for the relevant emotional category are listed in Tables 1 and 2 for EMO-DB and VAM, respectively. The *reference audio* is the number of audio samples used in the *reference set* for computing the emotional difference. The *reference audio* stands for the computed emotional differences between the *training* and the *reference* audio sets. The number of patterns in the *reference* under *training* is approximately the product of the *training* and the *reference* excluding the content belonging to the same person. For VAM class, q1 is specified as the reference set while it is switched to the emotional class *neutral* for EMO-DB corpus. Hence, 71 audio files from neutral mode in EMO-DB and 21 audio files from q1 category have been used as the reference set.

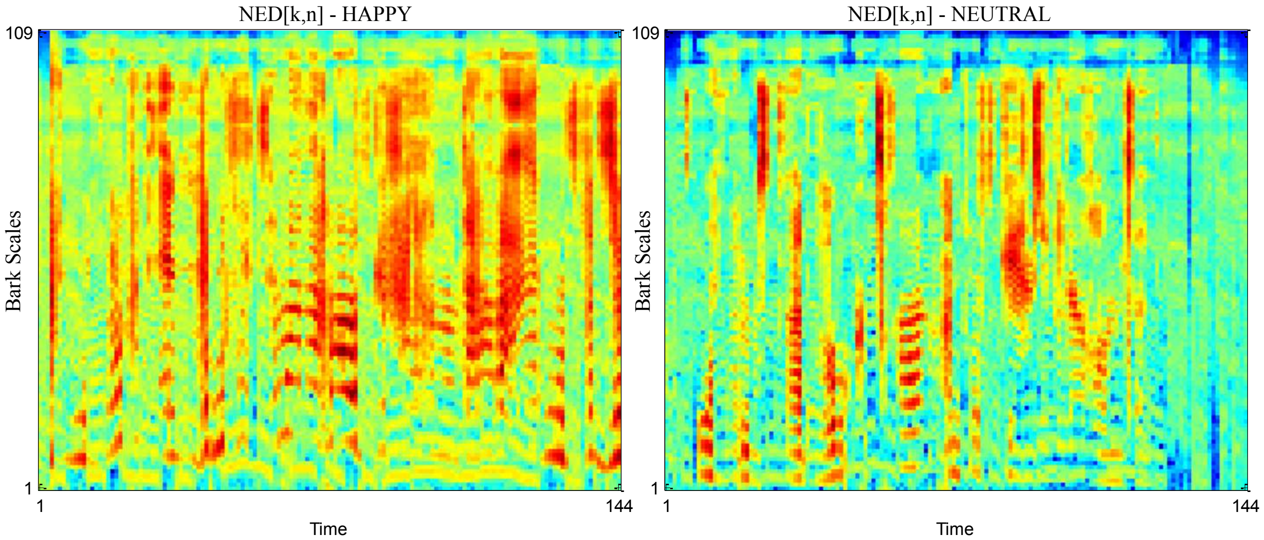


Figure 11 NED for happy and neutral samples taken from EMO-DB [15]. The higher-frequency components of happy and neutral become more discriminative with the help of NED in Bark scale.

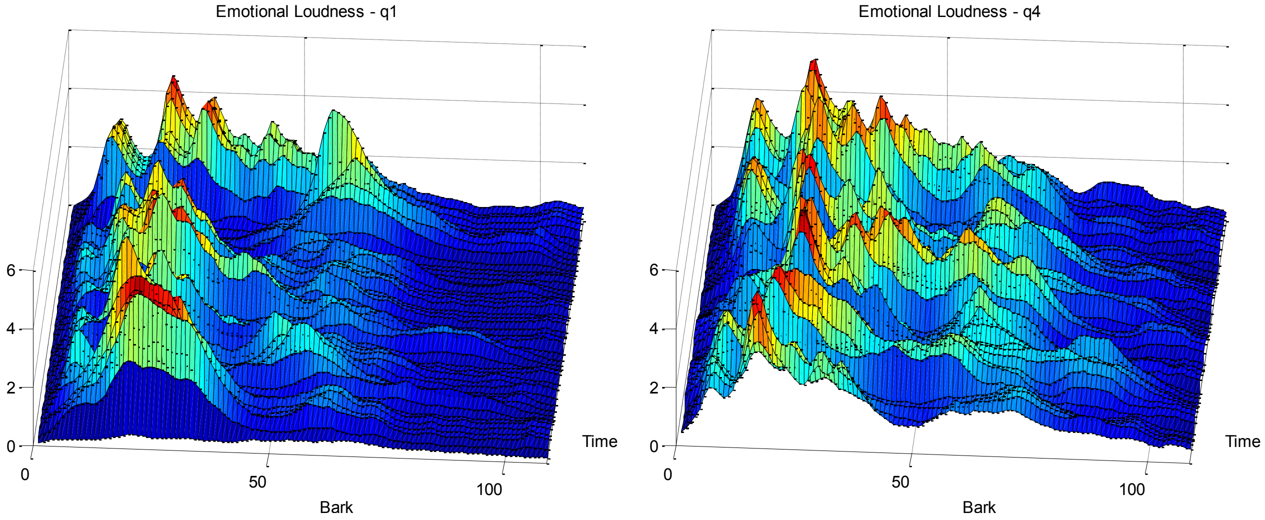


Figure 12 Loudness of the emotional data in time-Bark scale for q1 and q4. The figure illustrates the distribution of emotional loudness for q1 and q4 type speech data at each Bark scale through time. As expected, the loudness of q4, which is arousally more powerful, is dominant at higher frequencies, as loudness of q1 mainly remains at a lower frequency band.

In this study, we use a new decision rule, S-MV (explained in Section 4 with Equations 1-4) for the decision of the emotional class replacing the previous rule, majority voting (MV). Basically, the proposed system gives a decision on each observed feature vector based on its variations from the reference set; hence, the S-MV approach exploits all of the posterior probabilities assigned to each observed feature vector. Therefore, the

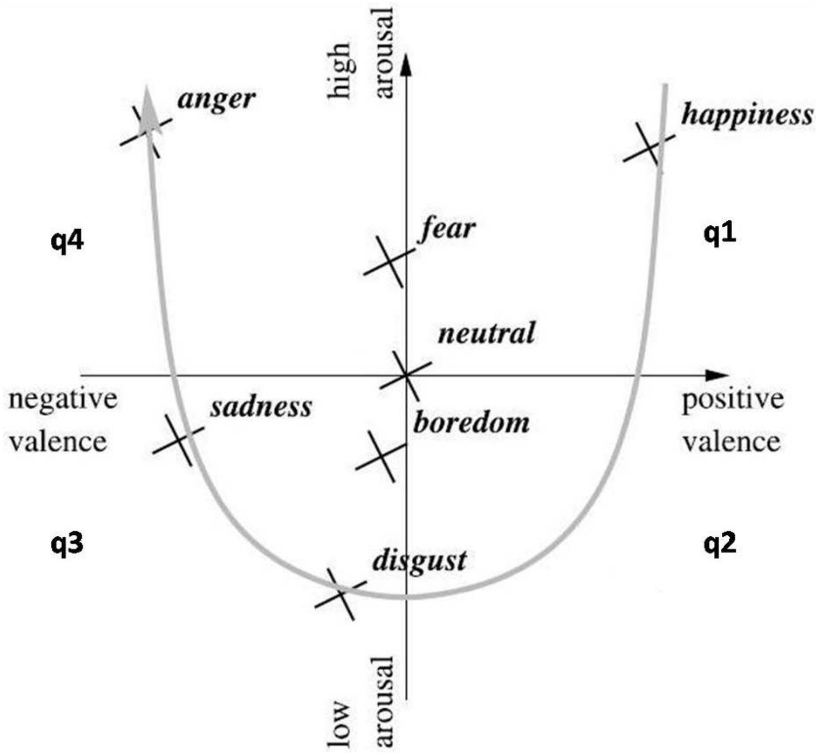


Figure 13 Two-dimensional emotion space mapping. The discrete emotions of EMO-DB are located on the arousal and valence axes. Regarding VAM, four quadrants are considered to evaluate the continuous emotion nature of the database on the arousal-valence axis. The mentioned quadrant's are assigned as "happy/excited" (q1), "angry/anxious" (q2), "sad/bored" (q3), and "relaxed/serene" (q4) [6,30].

Table 1 Number of training and test samples in EMO-DB [15]

Emotional classes	Training		Test	
	Number of feature vectors		Number of feature vectors	
	openEAR (6,556 features)	9-d perceptual feature vectors	openEAR (6,556 features)	9-d perceptual feature vectors
Angry	114	8128	13	902
Bored	73	4669	9	619
Happy	64	4546	7	502
Sad	56	3968	6	440
Neutral	70	5024	8	540
Disgusted	41	2917	5	447
Fear	90	6397	10	1264

number of vectors that needs to be considered for a decision is equal to the size of the reference set which enables to make a decision based on a broad statistical information about the data (recall that the size of reference set is 21 for VAM and 71 for EMO-DB). Taking greater amount of statistical data into consideration for a decision is the main advantage of the S-MV that makes it superior to MV.

At the classification stage, the *perceptual feature* vectors representing each test utterance are exposed to the classifiers, which have been trained prior to the classification. The posterior probabilities contributing to each reference set are evaluated for each category. The posterior probabilities for each reference are compared among categories and the category having the highest $P^i_{j\text{final}}, j = 1, 2$ score is decided for the relevant audio signal. As it is explained in Section 4, this enables us to assign the class label of the category which has the higher posterior with a smaller variance.

In the test cases, we applied Leave One Speaker Out (LOSO) and Leave One Speaker Group Out (LOSGO) strategies to evaluate speaker independency in EMO-DB and VAM, respectively. The main benefit of using LOGO and LOSGO methods is to be able to perform the test cases in a comparable environment with the state-of-the-art methods [5,9].

For our tests, we employ two types of classifiers for emotion detection, the GMM which has been implemented in MATLAB and the SVM using LibSVM [28] tool in WEKA [29]. We use GMM method because of its efficiency in modeling diverse statistics of the observed emotional categories. We also provide SVM to benchmark with openEAR [5]. Since LOSO and LOSGO methods are used and the utterances recorded in databases are not distributed equally between the speakers, the average number of training and test samples varies for each class. Tables 1 and 2, respectively, report the average number of the EMO-DB and VAM feature vectors used for the training and tests by the proposed system as well as the openEAR tool.

In the training phase, the 9-d feature vectors are extracted from the training patterns. To evaluate the recognition rate achieved using mixture models, the feature vectors are fed into the GMM/SVM classifier. Iterative expectation maximization algorithm is used to estimate the parameters of the mixture of Gaussian density functions representing each emotional class. The number of mixtures is determined as $j = 3$ on the basis of empirical evidences. It needs to be noted that the off diagonal components converge to small values when compared to the diagonal components; therefore, diagonal covariance matrices are employed at classification stage to reduce the computational complexity. Classification has been performed by using a Bayesian classifier with the help of S-MV. On the other hand, for the training of SVM classifier, a third-degree radial basis function has been used in the LibSVM tool with the parameters; "cost 100, gamma 10, loss 0.1 and nu 0.5". The number of support vectors learned by the SVM classifier is equal to approximately half of the training patterns.

Table 2 Number of training and test samples in VAM [16]

Emotional classes	Training		Test	
	Number of feature vectors		Number of feature vectors	
	openEAR (6,556 features)	9-d perceptual feature vectors	openEAR (6,556 features)	9-d perceptual feature vectors
q1	12	232	9	189
q2	39	819	11	231
q3	366	7686	85	1785
q4	350	7350	74	1554

There is an obvious imbalance between the perceptual features and conventional openEAR and HTK tools from the point of features and training test set sizes. As the perceptual model employs only 9 features, openEAR toolkit extracts 6,552 features as 39 functionals of 56 acoustic LLDs. A recent study [9] extracts 2D features from the reported 6,552 features with a GerDa. GerDa is denoted as a multilayer artificial neural network with many hidden layers and millions of free parameters learning discriminant features among a large set of acoustic features. High number of parameters may be a probable drawback of GerDa. On the other side in respect to low sized 9-d perceptual feature vectors, our *reference* set expands the size of the training and test sets proportional to the *reference* set size. This fact can be observable from the number of test and training samples reported in Tables 1 and 2.

Another key criterion that should be mentioned is that the audio files belonging to the test users are not included either in the training or the reference sets. This criterion is compliant with the practical cases where the test user has no data existing in the training or the reference sets. Similarly, while establishing the *reference* set, the emotional differences of the audio patterns belonging to different speakers are employed.

We have first evaluated the optimality of our feature set. In this study, we concentrate on the introduced nine perceptual features and we try to find the impact of these features on each emotional category by the help of sequential floating forward feature selection (SFFS) and forward feature selection (FFS). The feature subset selection methods are applied on both databases (VAM, EMO-DB). The evaluator of SFFS and FFS has been used to select the features having higher correlation with the emotion labels by the help of the 'CfsSubsetEval' option in WEKA. The SVM classifier is run for each category pairs in the databases and consequently an assumption is carried out for the best feature order. The affect of each feature on the performance can be seen in Figure 14.

As a result of feature subset evaluation, it is concluded that the nine perceptual features have different order of importance involving different databases hence all of them should be used to achieve high classification accuracy. It is also concluded that the required number of audio descriptors shows variations among categories. Contrary to the high arousal modes, emotional categories remaining at the positive side of the valence mode need to be classified with a higher number of features. On the other hand, the order of importance of these features is not the same for EMO-DB and VAM, probably because of the different content of acted and natural emotional corpus. However, the features AHSM, AEB, W_{E1}, and W_{E2} have the highest impact on classification. These features are derived from the perceptually masked "harmonic structure", "temporal distribution of excitation levels", and "perceptual bandwidth".

We have also evaluated the improvement achieved by the decision rule S-MV. In Table 3, the impact of the new decision rule (S-MV) is presented in comparison to

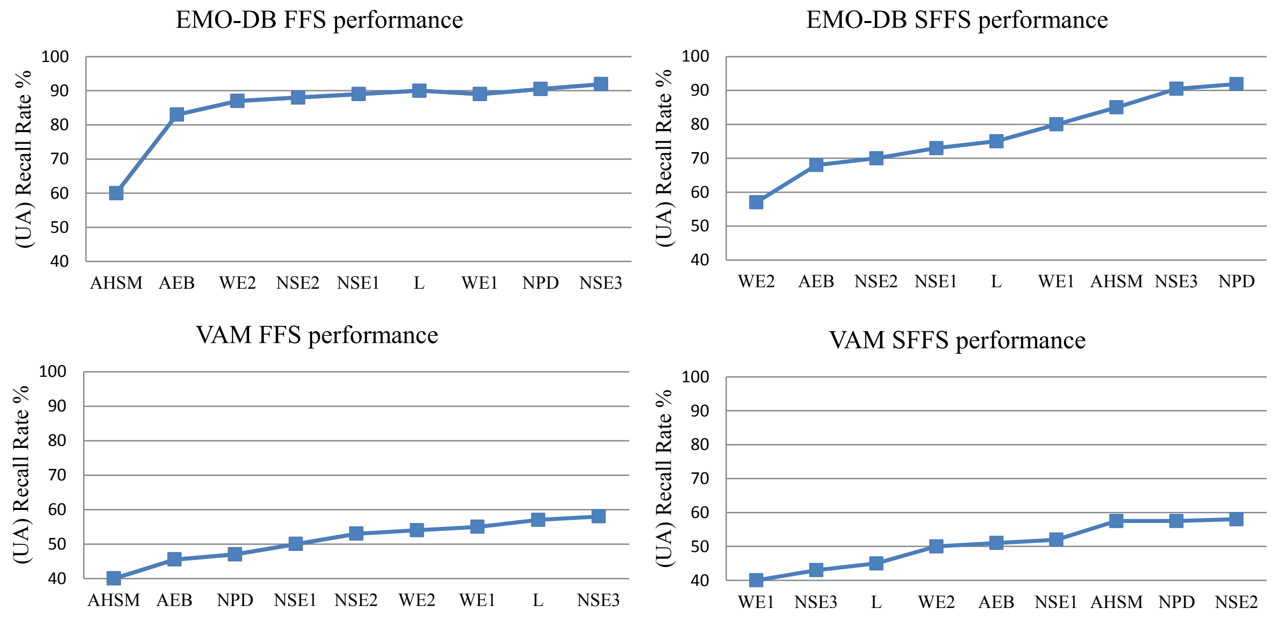


Figure 14 Outcomes of the FFS and SFFS evaluation performed on the VAM and EMO-DB. Recall rates achieved by the FFS and SFFS. Gain of the perceptual features is ordered from higher to lower impact on the recognition.

Table 3 Performance achieved by the perceptual features with soft majority voting in comparison to majority voting

		All (%)		Arousal (%)		Valence (%)	
		UA	WA	UA	WA	UA	WA
EMO-DB							
P-SVM	S-MV	86.3	85.9	95.2	95.1	94.3	95.6
	MV	85.2	86.1	91.3	92.7	88.0	88.1
VAM							
P-SVM	S-MV	61.3	76.2	69.4	71.7	59.9	83.3
	MV	57.1	74.7	59.2	73.3	52.4	77.8

MV where the superiority of the S-MV with SVM classifier can be observed. S-MV provides 1-6% improvement for EMO-DB and 4-10% improvement for VAM in all, arousal, and valence categories. It can be concluded that the S-MV outperforms the MV mainly because it enables us utilizing the statistical information stemming from the reference set effectively.

Finally, the audio emotion recognition rates achieved by the proposed perceptual features with various classifiers are evaluated for the binary arousal (passive, active) and valence (negative, positive) discrimination in VAM and EMO-DB databases. The results obtained by SVM and GMM classifiers using the perceptual features are, respectively, reported as P-SVM and P-GMM in Table 4. The emotion recognition rates provided by the state-of-the-art systems on the same databases are also reported in Table 4. The first generic impression that we get for all of the systems from Table 4 is that the performance reported for EMO-DB outperforms the recognition accuracy reached on VAM. This result is expected since the difficulty of perceiving and classifying natural spoken data is apparent. Another common point is all the systems in both databases provide higher performance in arousal with regard to valence. Unfortunately, there is no agreement within researchers on how acoustic features correlate with valence dimension. For example, both the *anger* and the *happiness* correspond to high arousal but they convey different affect which can be characterized by the valence dimension. A recent survey published in [2] confirms that the classification between high-activation (also called high-arousal) emotions and low-activation emotions can be achieved at high accuracies; however, classification between different emotions is still challenging. It is also reminded that conventional features are efficient only in distinguishing between high-arousal emotions, e.g., *anger*, *fear*, and *joy*, versus low-arousal ones, e.g., *sadness* [2] which results in inefficient performance for valence categories. These findings reveal the difficulty in detecting the valence mode.

Table 4 Emotion recognition rates achieved by the perceptual features and the state-of-the-art systems for EMO-DB and VAM corpus

	All (%)		Arousal (%)		Valence (%)	
	UA	WA	UA	WA	UA	WA
EMO-DB						
openEAR [8]	84.6	85.6	96.8	96.8	87.0	88.1
HTK [7]	73.2	77.1	91.5	91.5	78.0	80.4
GerDa [9]	79.1	81.9	97.6	97.4	82.2	87.5
P-SVM	86.3	85.9	95.2	95.1	94.3	95.6
P-GMM	92.5	90.4	92.1	90.0	91.9	95.2
VAM						
openEAR [8]	37.6	65.0	72.4	72.4	48.1	85.4
HTK [7]	38.4	70.2	76.5	76.5	49.2	89.9
GerDa [9]	39.3	68.0	78.4	77.1	52.4	92.3
P-SVM	61.3	76.2	69.4	71.7	59.9	83.3
P-GMM	60.2	74.1	66.1	70.8	58.0	69.1

A comparison of the P-SVM, and the P-GMM with regard to state-of-the-art methods reported in the literature such as openEAR, HTK, and GerDa can be made by looking at the emotion recognition rates reported in Table 4. P-SVM and P-GMM outperform other classifiers in all and valence tasks using the new decision rule S-MV. The P-SVM with S-MV provides a 7-16% improvement for EMO-DB and 7-11% in VAM for valence as the improvement of P-GMM is between 4-12% in EMO-DB and 6-10% in VAM for valence as well. The main reason for this might be that perceptual feature vectors model the valence axis more efficiently relative to classical features. Similar to openEAR, HTK, and GerDa, our perceptual features also uses energy and loudness; however, the major contribution comes from our perceptual model covering aspects such as spectral masking, outer and middle ear acoustic transform models, perceptual loudness, and perceptual bandwidth which all support the relatively more subjective position in valence. However, on the arousal axis GerDa plays a dominant role which is ahead of other classifiers (1-6% in EMO-DB, 2-12% in VAM). These results might indicate that GerDa may select compact and discriminative features with its layered structure for arousal. On the other hand, P-SVM and P-GMM express distinction in the overall. The advantage of perceptual features coming from valence detection might have an impact on the overall results as well.

7. Conclusions

In this article, we introduced a novel 9-d perceptual feature set for the task of acoustic emotion recognition. The proposed features show an improved recognition performance particularly for valence. The improvement in the performance is valid in both acted and natural emotions evaluated on EMO-DB and VAM corpus, respectively. We claim that this difference relies on the perceptual aspects of our method such as spectral masking, outer and middle ear acoustic transform models, besides the *reference* concept. Unlike the existing methods, the proposed perceptual feature vectors reflect the variation of every emotional audio sample from the "reference audio set". Our proposal is based on the hypothesis that the characteristic of the variations from any emotional mode is much more discriminative than the emotional data itself. This relatively objective criterion provides a positive contribution to emotion detection efforts, particularly for the distinction of positive and negative emotions.

We used our new feature set in a psychological emotion dimension (arousal-valence) model. The emotion recognition results over two popular databases illustrate a noticeable improvement over the previously reported baselines in "valence" and "all" categories. The P-SVM and P-GMM outperform other classifiers in "all" and "valence" tasks using the new decision rule S-MV. The P-SVM provides an improvement changing between 7-16% for EMO-DB and 7-11% in VAM for valence as the P-GMM improvement is between 4-12% range in EMO-DB and 6-10% range in VAM for valence. The main reason for this might be that perceptual feature vectors model the valence axis more efficiently relative to conventional features. However, on the arousal axis GerDa plays a dominant role which is ahead of (1-6% in EMO-DB, 2-12% in VAM) other classifiers. These results indicate that the nonlinear learning scheme of GerDa enables may select compact and discriminative features with its layered structure for arousal. The proposed P-SVM and P-GMM express distinction in the task "all". Higher recognition rates achieved for "all" mainly arise from the advantage of valence detection.

We have evaluated the impact of the perceptual features on each emotion category by the help of the FFS and the SFFS. As a result of feature subset selection, it is concluded that the impact of nine perceptual features are not the same on different emotional categories hence all of them should be used to improve the recognition accuracy. Despite of the high arousal modes, emotional categories remaining at the positive side of valence need to be classified with a higher number of features. On the other hand, the order of importance of these features is not the same for EMO-DB and VAM, probably because of the different content of acted and natural emotional corpus.

As the *reference* concept provides objectivity to the subjective nature of the emotion, the future study will pursuit the optimal reference set selection topic. The outstanding characteristic of our method relative to the conventional methods is the smaller number of features with an uncomplicated classifier. However, there is a computational complexity trade-off originating from the reference set size which is a result of *emotional difference* computation. This computational complexity is a vital issue to be investigated in the future study.

Abbreviations

AEB: average number of emotional blocks; AHSM: average harmonic structure of magnitude; DFT: discrete Fourier transform; EMO-DB: Berlin emotional speech database; F0: fundamental frequency; FFS: forward feature selection; GerDA: generalized discriminant analysis; GMM: Gaussian mixture model; HTK: hidden Markov toolkit; ITU: International Telecommunications Union; LLD: low level descriptor; LOSGO: leave one speaker group out; LOSO: leave one speaker out; MFCC: Mel-frequency cepstral coefficient; NPD: Normalized Perceptual Difference; NSE: normalized spectral envelope; openEAR: openEAR toolkit; PEAQ: perceptual evaluation of audio quality; q1-q4: quadrants in dimensional emotion; STFT: short-time Fourier transform; SVM: support vector machine; TEO: Teager energy operator; UA: unweighted average; VAM: Vera Am Mittag emotional database; WA: weighted average.

Competing interests

The authors declare that they have no competing interests.

Received: 11 November 2011 Accepted: 4 May 2012
Published: 4 May 2012

References

1. R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, J Taylor, Emotion recognition in human-computer interaction. IEEE Signal Process Mag. **18**(1), 32–80 (2001)
2. ME Ayadia, MS Kamelb, F Karrayb, Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011)
3. CM Lee, SS Narayanan, Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process. **13**, 293–303 (2005)
4. H Gunes, B Schuller, M Pantic, R Cowie, Emotion representation, analysis and synthesis in continuous space: a survey. in *Proc of the IEEE Int Workshop on EmoSPACE, in Conjunction with the IEEE FG 2011*, CA, USA 827–834 (March 2011)
5. B Schuller, B Vlasenko, F Eyben, G Rigoll, A Wendemuth, Acoustic emotion recognition: a benchmark comparison of performances. in *Proc of the IEEE Automatic Speech Recognition and Understanding Workshop*, Italy 552–557 (December 2009)
6. D Ververidis, C Kotropoulos, Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006)
7. S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book (v3.4)*, (Cambridge University Press, Cambridge, 2006)
8. F Eyben, M Wollmer, B Schuller, openEAR–introducing the munich open-source emotion and affect recognition toolkit. in *IEEE Proc of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction*, Amsterdam 576–581 (2009)
9. A Stuhlsatz, C Meyer, F Eyben, T Zielke, G Meier, B Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks. in *Proc of the IEEE International Conference on Acoustics Speech and Signal Processing,* Prague 5688–5691 (2011)
10. M Lugger, B Yang, Psychological motivated multi-stage emotion classification exploiting voice quality features, in *Speech Recognition, Technologies and Applications*, ed. by France Mihelic, Janez Zibert (I-Tech Education and Publishing, Vienna, Austria, 2008), pp. 395–410
11. B Yang, M Lugger, Emotion recognition from speech signals using new harmony features. Signal Process. **90**(5), 1415–1423 (2010)
12. HG Kim, N Moreau, T Sikora, *MPEG-7 Audio and Beyond*, (John Wiley & Sons Ltd., England, 2005)
13. C Sezgin, B Gunsel, GK Kurt, A novel perceptual feature set for audio emotion recognition. in *Proc of the IEEE Int Workshop on EmoSPACE, in Conjunction with the IEEE FG 2011*, CA, USA 780–785 (March 2011)
14. B Schuller, B Vlasenko, F Eyben, M Wollmer, A Stuhlsatz, A Wendemuth, G Rigoll, Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput. **1**(2), 1–13 (2010)

15. F Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, B Weiss, A database of German emotional speech. in *Proc of the INTERSPEECH*, Portugal 1517–1520 (2005)
16. M Grimm, K Kroschel, S Narayanan, The Vera am Mittag German audio-visual emotional speech database. in *Proc of the IEEE International Conference on Multimedia and Expo*, Germany 737–742 (2008)
17. P Ekman, An argument for basic emotions. Cognit Emotion. **6**, 169–200 (1992)
18. H Schlosberg, Three dimensions of emotions. Psychol Rev. **61**, 81–88 (1954)
19. JA Russell, A circumplex model of affect. J Personal Soc Psychol. **39**, 1161–1178 (1980)
20. T Nwe, S Foo, L De Silva, Speech emotion recognition using hidden Markov models. Speech Commun. **41**, 603–623 (2003)
21. G Zhou, JHL Hansen, JF Kaiser, Nonlinear feature based classification of speech under stress. IEEE Trans Speech Audio Process. **9**(3), 201–216 (2001)
22. L Chen, T Huang, T Miyasato, R Nakatsu, Multimodal human emotion/expression recognition. in *Proc of the IEEE Automatic Face and Gesture Recognition*, Japan 366–371 (April 1998)
23. P Pudil, F Ferri, J Novovicova, J Kittler, Floating search method for feature selection with nonmonotonic criterion functions. in *Proc of the International Conference on Pattern Recognition*, Israel 279–283 (October 1994)
24. International Telecommunications Union Recommendation BS.1387-1, Method for objective measurements of perceived audio quality (2000)
25. T Thiede, WC Treurniet, R Bitto, C Schmidmer, T Sporer, JG Beerends, C Colomes, M Keyhl, H Stoll, K Brandenburg, PEAQ–the ITU standard for objective measurement of perceived audio quality. J Audio Eng Soc. **48**, 3–29 (2000)
26. C Busso, S Lee, S Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans Audio Speech Lang Process. **17**(4), 582–596 (2009)
27. PJ Murphy, KG McGuigan, M Walsh, M Colreavy, Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. Acoust Soc Am. **123**(3), 1642–1652 (2008)
28. CC Chang, CJ Lin, LibSVM: a library for support vector machines. ACM Trans Intell Syst Technol. **2**, 27:1–27:27 (2001)
29. IH Witten, E Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*, (Morgan Kaufman, San Francisco, 2000)
30. E André, M Rehm, W Minker, D Bühler, Endowing spoken language dialogue systems with emotional intelligence. in *Proc of the Affective Dialogue Systems*, Germany 178–187 (June 2004)

doi:10.1186/1687-4722-2012-16
Cite this article as: Sezgin *et al.*: **Perceptual audio features for emotion detection.** *EURASIP Journal on Audio, Speech, and Music Processing* 2012 2012:16.