Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 1306 – 1324

doi:10.3906/elk-1312-196

How to categorize emotional speech signals with respect to the speaker's degree of emotional intensity

Salman KARIMI[1,2,*], **Mohammad Hossein SEDAAGHI**[1]
[1]Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran
[2]Department of Electrical Engineering, University of A. A. Boroujerdi, Boroojerd, Lorestan, Iran

Received: 24.12.2013 • **Accepted/Published Online:** 03.03.2014 • **Final Version:** 23.03.2016

Abstract: Recently, classifying different emotional content of speech signals automatically has become one of the most important comprehensive inquiries. The main subject in this field is related to the improvement of the correct classification rate (CCR) resulting from the proposed techniques. However, a literature review shows that there is no notable research on finding appropriate parameters that are related to the intensity of emotions. In this article, we investigate the proper features to be employed in the recognition of emotional speech utterances according to their intensities. In this manner, 4 emotional classes of the Berlin Emotional Speech database, happiness, anger, fear, and boredom, are evaluated in high and low intensity degrees. Utilizing different classifiers, a CCR of about 70% is obtained. Moreover, a 10-fold cross-validation procedure is used to enhance the consistency of the results.

Key words: Signal processing, paralinguistic parameters, emotional speech classification

1. Introduction

Nowadays, human emotion recognition plays a vital role in the field of human–computer interactions [1]. Different modalities can be used to identify affective moods of people, e.g., speech content, body language, facial expressions, and biological parameters [2]. Because of the simplicity of speech-related techniques and instruments, manifestation of emotional states via speech signals has become the most important topic among these domains. Therefore, many researchers all over the world do research on techniques to introduce new approaches in emotional speech recognition (ESR) methods. They have endeavored to improve the correct classification rate (CCR) of such systems in 3 major parts: feature extraction, feature selection, and classification.

Bezooijen [3] extracted one of the initial acoustic features for the purpose of emotional speech recognition. Using statistical properties of these features, he succeeded to classify emotions with low accuracies. Some researchers such as Tolkmitt and Scherer [4] followed his studies, but they did not achieve any noticeable success. Afterwards, McGilloway et al. extracted 32 prosodic features allied with pitch, tune, intensity, and spectrum of speech signals [5]. With the base of these features, some researchers, e.g., Hammal et al. [6], Ververidis et al. [7], Pao et al. [8], and Yang and Pu [9], extracted some additional features with respect to those introduced by McGilloway et al. Using different filter methods such as mutual information, entropy, and variation coefficient (VC)-based filters and some wrapper methods, e.g., sequential forward selection, sequential backward selection, sequential floating forward selection (SFFS), and sequential floating backward selection, they improved CCRs in normal situations.

*Correspondence: karimi.salman@gmail.com

In addition to these efforts, in order to prepare appropriate systems for being used in more real phenomena, some researchers worked on the topic of ESR in the presence of background noise [10–13]. The main purpose of these efforts was related to the robustness-improvement of ESR systems. Employing different preprocessing steps, they extracted suitable features to be applied for the classification of speech signals in accordance with their emotional content.

Literature reviews show that not enough research exists about the relations between the features and the intensity of emotional content of speech signals. The best reported work is that of Song et al. [14]. They only employed log frequency power coefficients as the main features for the assessment of emotional intensities in speech utterances.

In this paper we present a comprehensive study on the features that are related to the intensity of emotions in speech signals. To gain a superior understanding about these relations, we have focused on 4 emotions: anger and happiness, which are related to high arousal emotions, and fear and boredom as low arousal ones [15]. In this way, all of the evaluations of this paper are executed on the utterances of the Berlin Emotional Speech database (EmoDB), which were recorded in the room of the Technical Acoustics Department of the Technical University of Berlin [16].

This article is organized as follows: the proposed method of this paper and its different aspects are explained in Section 2. In Section 3, intensity-related features with respect to different emotions are extracted from speech signals. Subsequently, Section 4 discusses speech signals classified in accordance with their emotional contents, using the best selected features. In Section 5, by means of 2 classifiers and by using the best intensity-related features, emotional speech signals are categorized into 2 groups: high intensity level and low intensity level. Finally, conclusions are provided in Section 6.

2. Proposed method

Speech source and vocal tract are 2 factors that can be modified through the changes of human affects. Therefore, their evaluation could help us to recognize the emotional class of speakers and their intensities correctly. The cornerstone of the article is based on this fundamental concept.

The frameworks of the manuscript are illustrated in Figure 1. The first framework, illustrated in Figure 1a, is related to the process of classifying input speech signals to appropriate emotional categories. In the second framework, Figure 1b, the silent sections and noisy components are first removed from the utterances using the voice-activity detection (VAD) algorithm. Additionally, Figure 1c shows the total process of the paper.

Subsequently, in the feature extraction section, 284 features are extracted from high and low emotional intensity signals. Comparison of these features shows the effects of emotional levels on the speech samples. Afterwards, applying filter and wrapper methods, the best features are selected for use in the classification step of intensity levels where 3 classifiers, i.e. Bayes [17,18], linear, and Gaussian radial basis function-based kernel support vector machines (SVMs), are applied [19,20]. Those utterances of the EmoDB that have equal textual content and are spoken by 1 speaker in 1 emotional state are then divided into 2 classes with respect to their emotional-intensity level, high or low.

2.1. Preprocessing

Prior to the extraction of features, a VAD algorithm is applied to the utterances of the EmoDB as the preprocessing step. The VAD is used to solve the problem of separating active parts of speech signals from nonspeech sections [21].

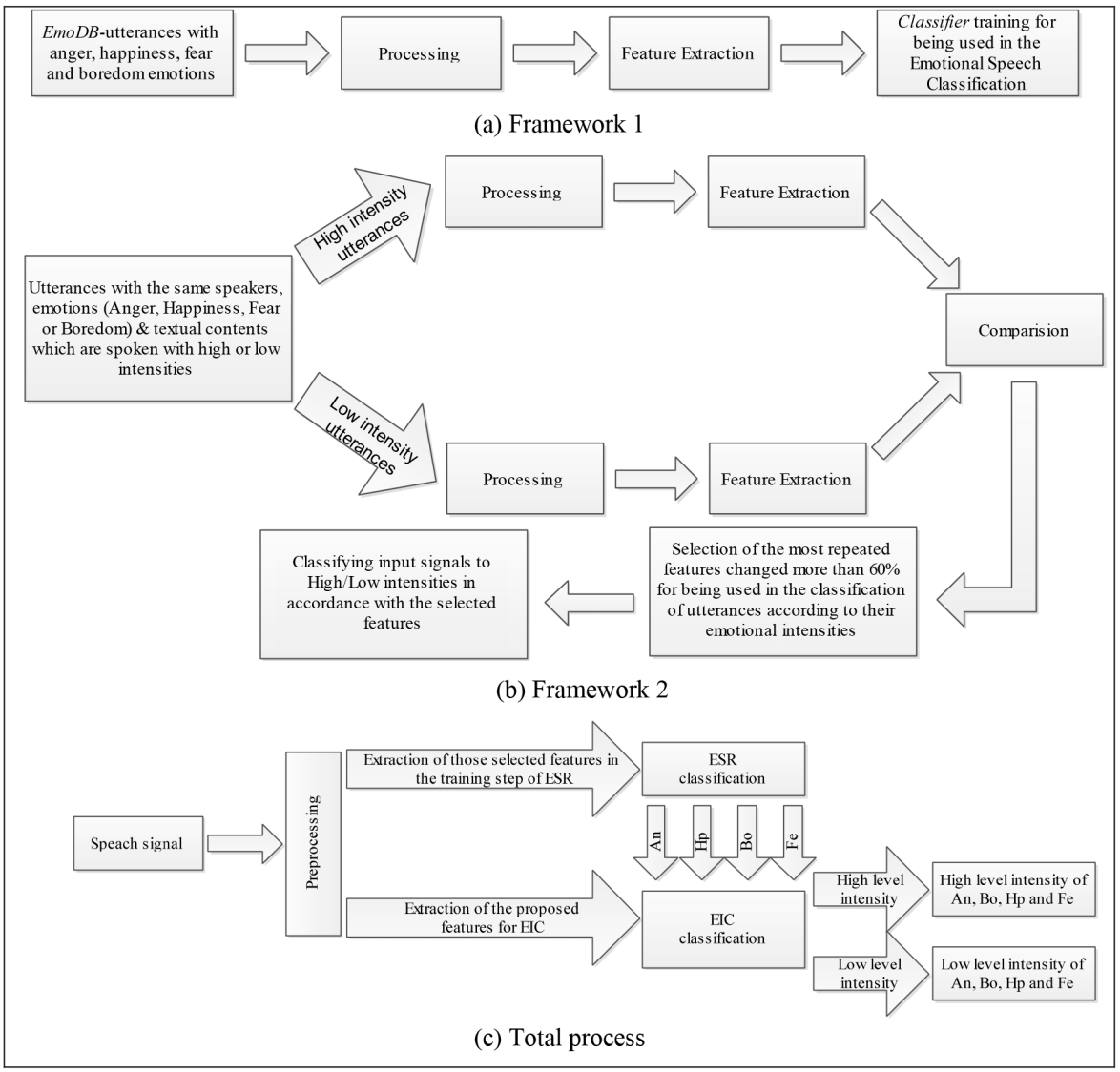

Figure 1. The frameworks utilized in the paper and sequential configuration of using them (ESR: emotional speech recognition; EIC: emotional intensity classification).

Literature reviews show that different VAD algorithms are available [22–24]. In this paper, a mixture of the short-time energy method and the zero crossing rate of speech signals is applied to locate active and silent parts of the utterances. In this way, a window with 15 ms length and 7.5 ms shift is employed to perform this algorithm.

2.2. Feature extraction

As mentioned before, this paper evaluates the intensity of emotional contents in speech-related utterances for the first time. In this way, we derived 284 features from the best known speech-related characteristics to improve the accuracy of the classification processes. As presented in Table 1, these features are related to 7 groups: formants, pitch, energy, spectrum, mel frequency cepstral coefficients (MFCCs), linear prediction coefficients (LPCs), and perceptual linear prediction (PLP) coefficients. Moreover, different functions used to extract the aforementioned features in this article are specified in Table 1. Most of them are obvious, except those related to

plateaus at minima and maxima in which we set the required thresholds to 75% of the minimum and maximum, respectively. The procedures of extracting these features are described as follows.

Table 1. Feature set.

Category (#)	Operators
Formants (16)	Mean, max, min, and variance of the first 4 formants.
Pitch (37)	Max, min, mean, median, interquartile range of pitch values. Pitch existence in the utterance expressed in percentage (0%–100%). Max, mean, median, interquartile of durations for the plateaus at minima. Mean, median, interquartile range of pitch values for the plateaus at minima. Max, mean, median, interquartile range, upper limit (90%) of durations for the plateaus at maxima. Mean, median, interquartile of the pitch values within the plateaus at maxima. Max, mean, median, interquartile range of durations of the rising slopes. Mean, median, interquartile of the pitch values within the rising slopes. Max, mean, median, interquartile duration of the falling slopes. Mean, median, interquartile of the pitch values within the falling slopes.
Energy (34)	Max, min, mean, median, interquartile range of energy values. Max, mean, median, interquartile range of durations for the plateaus at minima. Mean, median, interquartile range of energy values for the plateaus at minima. Max, mean, median, interquartile range, upper limit (90%) of duration for the plateaus at maxima. Mean, median, interquartile range of the energy values within the plateaus at maxima. Max, mean, median, interquartile range of durations of the rising slopes of energy contours. Mean, median, interquartile range of the energy values within the rising slopes of energy contours. Max, mean, median, interquartile range of durations of the falling slopes of energy contours. Mean, median, interquartile range of the energy values within the falling slopes of energy contours.
Spectral (43)	Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, 3950 Hz. Energy in the frequency bands 250–600, 600–1000, 1000–1500, 1500–2100, 2100–2800, 2800–3500, 3500–3950 Hz. Features (101– 106): energy in the frequency bands 250–1000, 600–1500, 1000–2100, 1500–2800, 2100–3500, 2800–3950 Hz. Features (107–111): energy in the frequency bands 250–1500, 600–2100, 1000–2800, 1500–3500, 2100–3950 Hz. Features (112–113): energy ratio between the frequency bands (3950–2100) and (2100–0) and between the frequency bands (2100–1000) and (1000–0). Energy in the frequency bands 250–2100, 600–2800, 1000–3500, 1500–3950, 250–2800, 600–3500, 1000–3950, 250–3500, 600–3950, and 250–3950 Hz. Energy ratio between the frequency bands (3950–3500) and (3500–0), (3950–2800) and (2800–0), (3950–1500) and (1500–0), (3950–600) and (600–0), (3950–250) and (250–0).
MFCCs (52)	Mean, max, min, and variance.
LPCs (52)	Mean, max, min, and variance.
PLPs (52)	Mean, max, min and variance.

In the speech-related sciences, the resonances of the vocal tract are called formants, which are described with 2 important parameters: formants' locations and their bandwidths [25].

These parameters are calculated using a transformation from complex root pairs $\zeta = \alpha e^{\pm\Phi}$ to formant frequency F and 3 dB bandwidths B as follows [25–27]:

$$F = \frac{f_s}{2\pi}\Phi, \tag{1}$$

$$B = -\frac{f_s}{\pi}\ln\alpha, \tag{2}$$

where f_s stands for the sampling frequency (SF). In order to approximate the time periods of different formants, LPCs should be correctly calculated in accordance with the SF of sounds. Utterances of the EmoDB are recorded with SF = 16 kHz, so, in order to improve the accuracy of the extracted 3rd and 4th formants, the order of

LPCs is set to 18 [28]. Additionally, the first 4 formants are computed in consecutive 20 ms length frames, which have 50% overlap with the previous ones. Four statistical parameters of these formants have been estimated among all the active frames. The mathematical formulation of these parameters, which are mean, max, min, and variance, are explained in Eqs. (3)–(6).

$$MEANF1 = \sum_{i=1}^{n} F1_i \Big/ n \tag{3}$$

$$MAXF1 = \max_{i=1}^{n} (F1_i) \tag{4}$$

$$MINF1 = \min_{i=1}^{n} (F1_i) \tag{5}$$

$$VARF1 = \sum_{i=1}^{n} (F1_i - MEANF1)^2 \Big/ (n-1) \tag{6}$$

Here, $F1_i$ is the value of the first formant of the ith active frame of the speech signal and n is the number of active frames in each utterance. Other formant-related features (FRFs) are extracted by similar rules as in the above equations while $F2$, $F3$, and $F4$ are employed instead of $F1$. FRFs are located in the first 16 indices of our feature set.

Table 1 shows that the second group of features utilized in this article are related to pitch frequencies [29]. Pitch periods (T), which are formed with the fluctuations of vocal folds, are connected with 2 biological parameters, barometric pressure in the subglottal cavity and the tension of the vocal cords. Having similar characteristics to the studies of Sondhi and Ververidis and Kotropoulos [30,31], a technique established on the autocorrelation of a center-clipped frames procedure is applied in this article. Thus, speech signals are passed from a low-pass filter with a cut-off frequency of 900 Hz. In order to improve the precision of the results, a 20 ms Hanning window, which is expressed in Eq. (7), is used as a windowing procedure to segment speech signals of the EmoDB such that their SF is 16 kHz.

$$\vartheta(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{320}\right)\right), 0 \le n \le 319 \tag{7}$$

Using this procedure, speech signals are segmented to $S_H(\mu, \xi)$ units where $\mu$and ξ are the midpoint and length of the segment, respectively. Now pitch frequencies can be estimated in these segments using Eq. (8).

$$F_0 = \frac{f_s}{N_w} \arg\max_{\beta} \left(|\Upsilon(\beta, \xi)|\right) \Big| \begin{array}{l} \beta = N_w (f_{\max}/f_s) \\ \beta = N_w (f_{\min}/f_s) \end{array} \tag{8}$$

In Eq. (8), f_s is equal to 16 kHz, f_{min} and f_{max} are the lowest and highest audible frequencies, and $\Upsilon(\beta, \xi)$ is the autocorrelation of speech segments and can be approximated as in Eq. (9).

$$\Upsilon(\beta, \xi) = \frac{1}{N_w} \sum_{c=\xi - N_w + 1}^{\xi} S_H(\mu, \xi) S_H(c - \beta, \xi) \tag{9}$$

In conformity with the segment number, extracted pitch values are concatenated in row vectors with respect to the signal numbers. All of the functions stated in the pitch-related features in Table 1 are then executed on the extracted contours.

The third section of the extracted features is related to the energy of speech signals. In this way, using Eq. (10), short-term energy values are measured for the aforementioned 20 ms length segments of speech signals.

$$\Omega(\alpha) = \sum_{\kappa=(\alpha-1)\xi+1}^{\alpha\xi} \left(S_H(\alpha,\kappa)\right)^2; \quad \alpha = 1,2,...,\eta \tag{10}$$

Here, $S_H(\alpha,\kappa)$ stands for the κth sample of the αth segment of the speech signal. Additionally, ξ and η are related to the length and number of the frames, respectively. Then Ω, which can be nominated as the energy vector of the speech signal, is utilized by the energy-related functions of Table 1, which are specified by indices 52–85 in our feature set.

Features indexed from 86 to 128 in the available feature set are related to the spectral content in certain frequency bands. In order to extract suitable features for application in telephony applications, we extracted these features from frequencies below 3950 Hz.

The next group of this feature set is related to MFCCs. Using MFCCs could enable researchers to classify different linear and nonlinear properties of speech signals [32]. After windowing each of the signals, the first 13 MFCCs are extracted from each of the segments. The mean, max, min, and variance of each of the MFCCs extracted from successive frames are then calculated to generate features indexed from 129 to 180.

LPC-related components make the other group of features indexed from 181 to 232. They supply precise approximations of a variety of speech-related properties. These coefficients are calculated using an orthogonal covariance method like that of Ning and Whiting [33]. In this way, the 1st to 13th LPCs are computed in all of the 20 ms length segments of each speech signal. The mean, maximum, minimum, and variance of each order of these coefficients are then computed between all windowed segments.

The last 52 features used in this article are related to the properties of PLPs [34]. In this way, the relative spectral (RASTA) PLP [35], which has more robust factors against different spectral distortions, is utilized as a substitute for pure PLP. Similar to MFCC- and LPC-related features, RASTA PLP-connected features have been extracted from successive windowed segments of speech signals with 20 ms length and 10 ms shift. In this manner, the mean, maximum, minimum, and variance of the first 13 PLP coefficients have been calculated and made into features indexed from 233 to 284 in Table 1.

3. Intensity-related features

In this section we evaluate those utterances that are related to a sentence spoken by 1 speaker in 1 feeling but with different intensities.

3.1. Anger-related utterances

Some of the statements of the EmoDB, e.g. "Das will sie am Mittwoch abgeben" ("She will hand it in on Wednesday") [16], have been expressed several times by 1 person in an angry mood, such as *03a02Wb* and *03a02Wc*. The differences of these utterances are related to the different intensity levels of expressed anger. The variation percentage of each feature among such utterances that have equal textual and emotional content

and are spoken by only 1 person is calculated as expressed in Eq. (11).

$$AFV = \frac{\left|\Psi_m^\lambda - \Psi_n^\lambda\right|}{\max\left(\Psi_m^\lambda - \Psi_n^\lambda\right)};\ \begin{cases} \lambda = 1, \ldots, 7 \\ m, n = 1, 2, \ldots, 535 \end{cases} \tag{11}$$

Here, Ψ indicates the feature-matrix extracted from all of the utterances of the EmoDB such that its columns and rows are related to features and utterances, respectively. Additionally, λ takes values from 1 to 7, which are related to anger, neutral, fear, boredom, happiness, sadness, and disgust, respectively. Moreover, m and n specify target rows of Ψ. For example, vectors Ψ_2^1 and Ψ_3^1 refer to the features of the second and third anger-related (first emotion) utterances. In this equation, AFV stands for absolute feature variation. In order to have a better understanding of the influences of emotional intensities on the behavior of the speech signals, Figure 2 demonstrates differences of such signals.

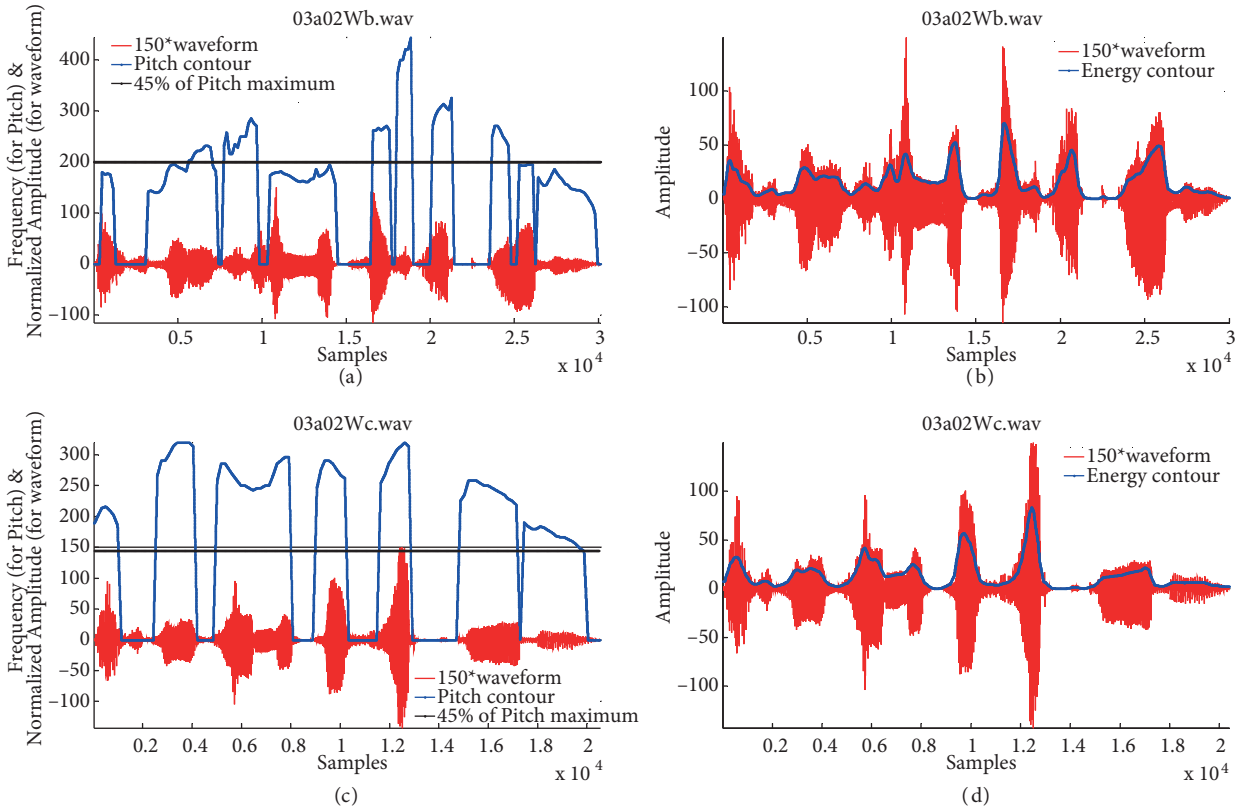

Figure 2. Pitch and energy contour illustration of the second statement of the EmoDB when it is spoken by the first speaker in the angry mood: a, b) low intensity; c, d) high intensity.

Additionally, using Eq. (11), Figure 3 illustrates the percentage of variations of each feature in the fifth and sixth anger-related utterances of the EmoDB (03a05Wa.wav and 03a05Wa.wav), which are spoken by the first speaker. We repeated this procedure for all the utterances expressed in an angry mood by 10 speakers of the EmoDB altering their intensity of expressing anger. Summarized results of these evaluations are reported in Table 2. As can be seen, in the 2nd column of this table, features that are changed by less than 3% are mentioned while those features which are changed by more than 60% are shown in column 3. This table shows

that, most of the time, features indexed by 25, 26, 29, 34, 41, 48, 64, 68, 71, 86, and 87 are altered by more than 60% while the intensity of angry mood of speech sounds is changing.

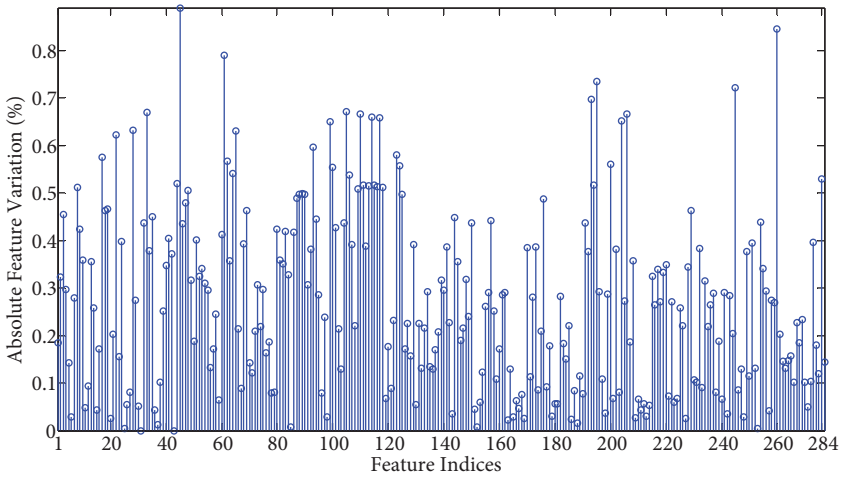

Figure 3. Indication of feature changes when speaker 1 in the EmoDB expresses second statement of this database in an angry mode with high and low intensities.

Table 2. Features that are modified less than 3% or more than 60% when a speaker alters his intensity of expressing anger in the EmoDB. NEDAA: Not enough data available for analysis.

Speaker index	Most repeated features changed less than 3%	Most repeated features changed more than 60%
1	32, 38, 45, 88, 96, 110, 118, 122, 124, 166, 213, 219, 255, 263	13, 25, 26, 29, 34, 41, 48, 63, 64, 68, 71, 87, 128, 198
2	-	23, 25, 26, 29, 34, 64, 48, 86, 87, 165, 166, 180
3	25, 45, 96, 103, 109, 110, 114, 115, 122, 124, 156, 166, 255	29, 39, 41, 48, 64, 68, 71, 86, 87, 94, 128, 163, 167, 180, 259, 263, 283
4	NEDAA	NEDAA
5	NEDAA	NEDAA
6	NEDAA	NEDAA
7	-	24, 25, 26, 29, 34, 41, 48, 61, 62, 63, 64, 65, 68, 71, 86, 87, 179, 248
8	32, 38, 45, 50, 96, 103, 105, 106, 109, 110, 114, 115, 118, 122, 124, 127, 133, 162, 166, 176, 211, 224, 255, 264, 269	9, 23, 24, 25, 26, 27, 28, 29, 34, 38, 39, 41, 48, 53, 64, 67, 68, 71, 86, 87, 141, 148, 167, 175, 176, 180, 186, 229, 254
9	4, 5, 12, 16, 22, 32, 34, 38, 45, 57, 85, 96, 100, 103, 109, 114, 115, 118, 122, 124, 136, 166, 167, 183, 185, 191, 215, 217, 221, 247, 251, 255, 268	8, 18, 25, 26, 29, 34, 41, 48, 53, 59, 64, 68, 71, 86, 155, 197, 259
10	32, 38, 45, 103, 109, 110, 114, 115, 122, 124, 125, 135, 139, 145, 160, 168, 250, 255, 272, 282	9, 18, 23, 25, 26, 27, 28, 29, 34, 41, 48, 53, 54, 55, 56, 59, 61, 62, 63, 64, 68, 71, 76, 77, 78, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 101, 102, 107, 112, 116, 119, 121, 130, 143, 166, 175, 234, 247, 254, 259, 260, 26

3.2. Happiness-related utterances

Some of the statements of the EmoDB, such as "An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht" ("At the weekends I have always gone home and seen Agnes"), are spoken in different

intensities of happiness by each of the speakers (e.g., 08b01Fd.wav and 08b01Fe.wav). Figure 4 shows some of the results of different emotional intensities on the structure of speech parameters.

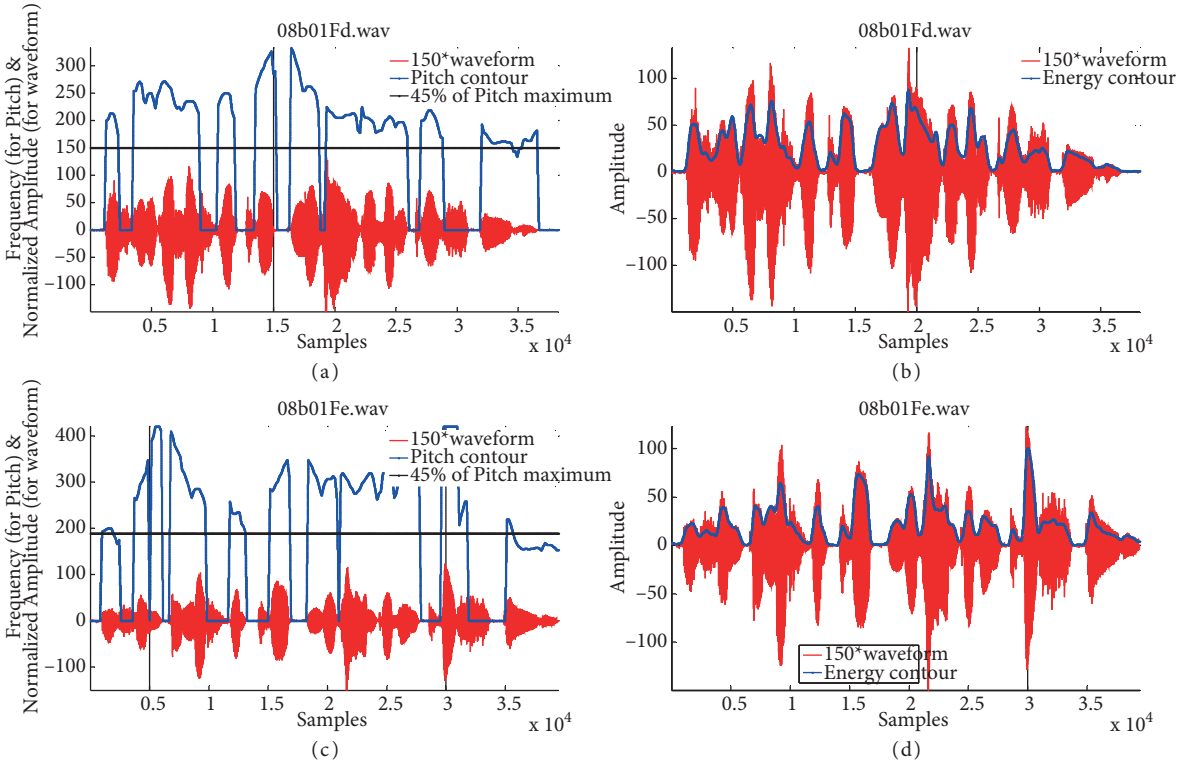

Figure 4. Pitch and energy contour illustration of the sixth statement of the EmoDB when it is spoken by the second speaker in a happy mood: a, b) low intensity; c, d) high intensity.

By means of Eq. (11), percentages of variations of each feature in those utterances that have equal speaker and textual content and are spoken in a happy mood with different intensities are evaluated. Results of these assessments are reported in the Table 3.

Evaluations of this table show that features represented by indices of 18, 24, 25, 26, 29, 34, 41, 48, 62, 64, 65, 68, 71, 86, 165, and 178 are the most repeated features changed by more than 60%. Comparison between these features and those related to the intensity of angry mood shows a good match so that features indexed as follows are changed by more than 60% in both of these feelings in accordance with different emotional intensities:

- Features 25 and 26: Median and interquartile range of pitch duration for the plateaus at minima, respectively.
- Feature 29: Interquartile range of pitch value for the plateaus at minima.
- Feature 34: Upper limit (90%) of pitch duration for the plateaus at maxima.
- Feature 41: Interquartile range of pitch duration of the rising slopes of pitch contours.
- Feature 48: Interquartile range of pitch duration of the falling slopes of pitch contours.
- Features 64 and 68: Maximum and upper limit (90%) of energy duration for the plateaus at maxima.

Table 3. Features with less than 3% change and those with more than 60% adjustment when a speaker alters his intensity of expressing happiness in the EmoDB.

Speaker index	Repeated features changed less than 3%	Repeated features changed more than 60%
1	9, 22, 27, 28, 45, 46, 57, 110, 115, 130, 138, 143, 157, 158, 165, 166, 172, 177, 185, 188, 217, 243, 261, 268, 277, 281	18, 24, 25, 29, 34, 39, 41, 44, 48, 59, 62, 65, 68, 71, 75, 79, 86, 160, 165, 174, 178, 197, 199, 283
2	13, 22, 45, 110, 115, 136, 143, 144, 146, 159, 160, 170, 172, 184, 189, 217, 240, 282	9, 23, 24, 25, 26, 29, 32, 34, 46, 48, 58, 60, 64, 65, 67, 68, 71, 72, 86, 128, 165, 174, 178, 179, 180, 205, 228, 231, 251, 258, 263, 269, 278, 281, 284
3	3, 22, 32, 45, 57, 103, 104, 105, 109, 110, 111, 114, 115, 118, 136, 138, 160, 172, 179, 184, 196, 210, 217, 221, 223, 240	8, 14, 16, 18, 24, 25, 26, 29, 38, 39, 41, 48, 56, 61, 62, 64, 65, 67, 68, 71, 80, 84, 85, 86, 87, 88, 94, 101, 141, 154, 165, 178, 199, 241, 242, 251, 252, 254, 267, 270, 279
4	NEDAA	NEDAA
5	15, 16, 22, 38, 41, 45, 57, 110, 112, 115, 119, 121, 136, 138, 139, 141, 159, 160, 165, 172, 184, 185, 209, 210, 217, 222, 234, 239, 240, 267, 275, 282	9, 18, 24, 25, 26, 27, 29, 34, 39, 41, 48, 62, 64, 65, 68, 71, 85, 86, 156, 160, 165, 174, 178, 201, 231
6	NEDAA	NEDAA
7	32, 36, 45, 51, 57, 61, 99, 110, 115, 136, 137, 138, 144, 160, 162, 172, 173, 184, 192, 196, 197, 217, 240, 274, 278	9, 18, 23, 24, 25, 26, 29, 34, 39, 41, 44, 48, 51, 62, 64, 65, 66, 67, 68, 71, 79, 86, 164, 165, 166, 167, 178, 179, 180, 200, 201, 227, 266, 280, 281
8	4, 22, 26, 45, 57, 69, 75, 96, 97, 100, 110, 115, 129, 130, 133, 136, 138, 144, 155, 160, 162, 165, 169, 172, 176, 184, 205, 211, 214, 215, 217, 224, 233, 240, 246, 247, 273, 275	18, 23, 24, 25, 26, 27, 28, 29, 33, 34, 39, 41, 44, 48, 51, 60, 62, 63, 64, 65, 66, 67, 68, 68, 71, 86, 128, 165, 167, 173, 175, 178, 179, 180, 189, 189, 199, 202, 202, 252, 271, 283
9	3, 4, 5, 22, 45, 49, 57, 110, 115, 126, 136, 138, 157, 160, 164, 172, 211, 213, 214, 217, 222, 240, 261, 263, 264	18, 24, 25, 29, 34, 38, 39, 41, 46, 48, 56, 59, 62, 63, 64, 65, 67, 68, 71, 72, 78, 88, 89, 90, 91, 92, 93, 95, 101, 102, 107, 108, 112, 113, 116, 117, 119, 121, 128, 165, 175, 178, 252
10	11, 20, 22, 35, 37, 45, 50, 57, 60, 67, 68, 82, 98, 103, 105, 109, 110, 111, 115, 123, 124, 126, 127, 128, 136, 138, 153, 160, 163, 172, 183, 185, 192, 193, 206, 217, 225, 232, 239, 240, 253, 270, 279, 280	18, 25, 26, 27, 29, 34, 41, 44, 46, 48, 51, 52, 53, 54, 56, 62, 63, 64, 65, 68, 71, 72, 73, 76, 78, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 119, 121, 139, 140, 165, 166, 177, 178, 179, 190, 255, 260, 261, 265, 280

- Feature 71: Interquartile range of the energy value within the plateaus at maxima.
- Feature 86: Energy below 250 Hz respectively.

As illustrated in Figure 5, according to the Geneva emotional wheel [15], anger and happiness are related to high control emotions. In order to extend the results of the obtained intensity-related features to a wider range of emotions, evaluations are also implemented on the extracted features of fear and boredom moods, which are related to low control ones.

3.3. Fear-related utterances

Compared to angry and joyful utterances, the EmoDB contains a number of fear-related sounds with similar textual and emotional content spoken by a speaker with different intensities. In order to have a better understanding of the differences of these signals, energy and pitch contours of such signals are illustrated in Figure 6. Additionally, Table 4 reports the variations of features through the use of Eq. (11).

Based on Table 4, features described by indices 9, 18, 23, 26, 29, 39, 48, 57, 58, 64, 71, and 204 are the most repeated features changed by more than 60%.

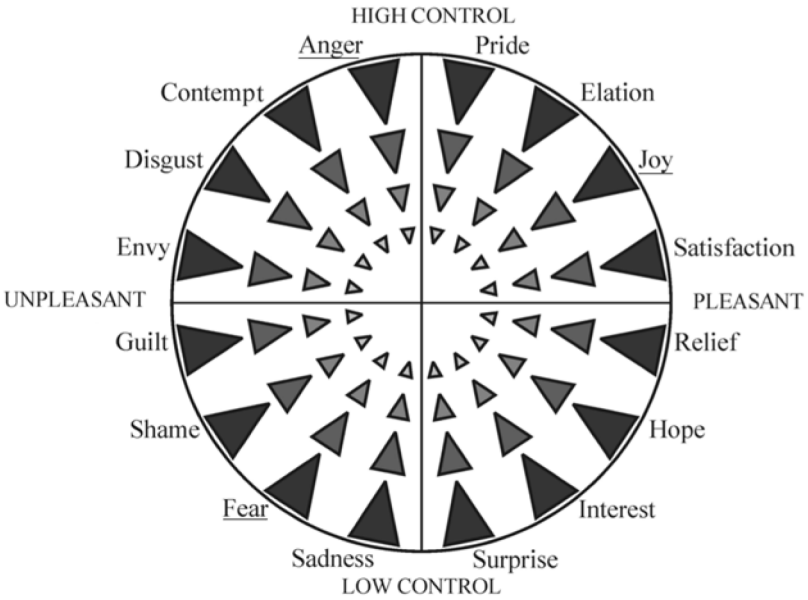

Figure 5. Geneva emotional wheel.

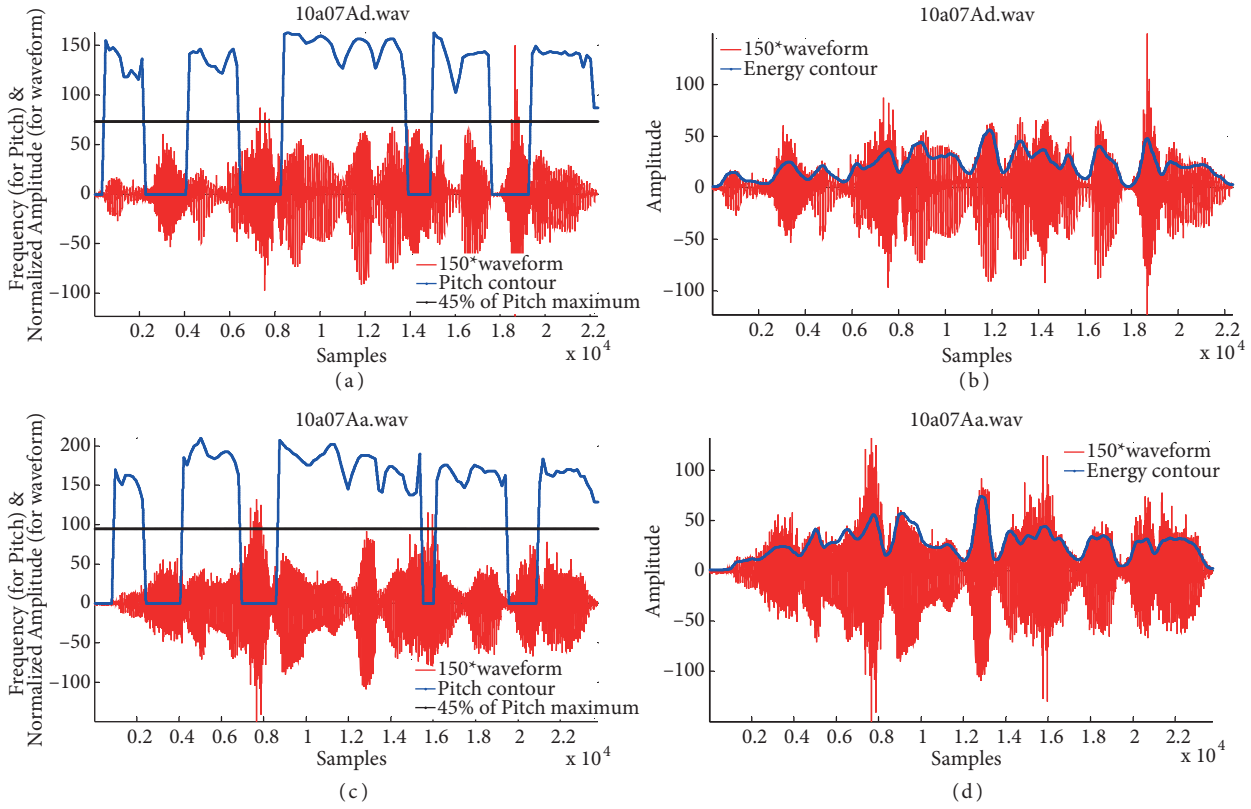

Figure 6. Pitch and energy contour figures of the fifth statement of the EmoDB when it is spoken by the second speaker in fear: a, b) low intensity; c, d) high intensity.

Table 4. Indices of features that are altered less than 3% or more than 60% when a speaker alters his intensity of expressing fear in the EmoDB.

Speaker index	Repeated features changed less than 3%	Repeated features changed more than 60%
2	19, 20, 30, 36, 37, 45, 60, 96, 100, 110, 115, 136, 158, 191, 193, 211, 217, 219, 222, 237, 251, 271	9, 18, 23, 25, 26, 29, 34, 38, 39, 48, 56, 57, 58, 59, 64, 65, 66, 67, 68, 71, 84, 85, 88, 89, 90, 91, 102, 107, 108, 112, 117, 119, 120, 121, 123, 157, 173, 175, 178, 204, 206
4	11, 32, 45, 53, 96, 110, 115, 130, 136, 157, 166, 189, 212, 215, 217, 235, 236, 237, 248, 265, 267, 282	8, 9, 18, 22, 23, 26, 29, 34, 38, 39, 42, 43, 48, 49, 57, 58, 64, 70, 71, 73, 82, 130, 143, 150, 168, 174, 180, 190, 199, 202, 203, 204, 245, 247, 251, 276, 277
5	12, 27, 45, 54, 55, 73, 74, 96, 97, 99, 105, 110, 115, 125, 130, 136, 138, 139, 144, 150, 155, 157, 160, 171, 175, 180, 182, 183, 184, 209, 214, 215, 217, 224, 226, 236, 237, 240, 243, 252, 281	9, 18, 26, 28, 29, 30, 33, 39, 41, 44, 45, 46, 46, 48, 57, 58, 60, 71, 81, 153, 178, 197, 201, 204, 206, 252
8	6, 10, 23, 25, 34, 34, 39, 41, 45, 51, 61, 65, 75, 76, 89, 96, 100, 110, 117, 134, 135, 136, 138, 142, 143, 153, 160, 164, 172, 196, 213, 217, 236, 237, 246, 249, 269	2, 3, 5, 9, 18, 21, 23, 25, 26, 29, 37, 38, 39, 45, 46, 48, 57, 58, 59, 60, 63, 64, 68, 71, 72, 86, 86, 97, 102, 108, 113, 117, 120, 126, 128, 150, 168, 175, 179, 201, 202, 204, 236, 243, 244, 258, 272, 283, 284
9	3, 12, 43, 45, 50, 96, 110, 115, 136, 140, 141, 166, 184, 195, 217, 236, 237, 246, 253, 278	9, 23, 29, 39, 46, 48, 49, 57, 58, 63, 64, 67, 71, 108, 113, 120, 153, 201, 204, 230, 257

3.4. Boredom-related utterances

According to the 3 aforementioned emotional categories, intensity-related features of boredom speech signals have been studied among the utterances of the EmoDB. In this way, some of the effects of these intensity differences on the pitch and energy parameters of speech signals are illustrated in Figure 7. Table 5 shows the most repeated features that are modified by less than 3% and more than 60%. Features explained in the third column of this table are highly connected with the boredom intensity of signals.

Table 5. Indices of those features that are modified less than 3% or more than 60% when the tenth speaker of the EmoDB changes her intensity of expressing boredom.

Indices of boredom-related signals; (related to 10th speaker)	Most repeated features that are changed less than 3%	Most repeated features that are changed more than 60%
70, 71	12, 45, 131, 140, 167, 235, 250, 262, 263, 264	9, 23, 24, 25, 26, 29, 32, 48, 55, 58, 59, 61, 64, 71, 77, 84, 96, 148, 198, 201, 204, 242, 279
73, 74		
75, 76		
79, 80		

Comparison between the most sensitive features related to the intensity of fear and boredom in speech sounds shows that those indexed by 9, 23, 26, 29, 48, 58, 64, 71, 201, and 204 are common in both of these emotions. These features are as follows:

- Feature 9: Minimum value of the first formant in the utterance.
- Feature 23: Maximum, mean, median, and interquartile range of duration for the plateaus at minima.
- Feature 26: Median and interquartile range of pitch duration for the plateaus at minima, respectively.
- Feature 29: Interquartile range of pitch value for the plateaus at minima.
- Feature 48: Interquartile of the duration range of the falling slopes of pitch contours.

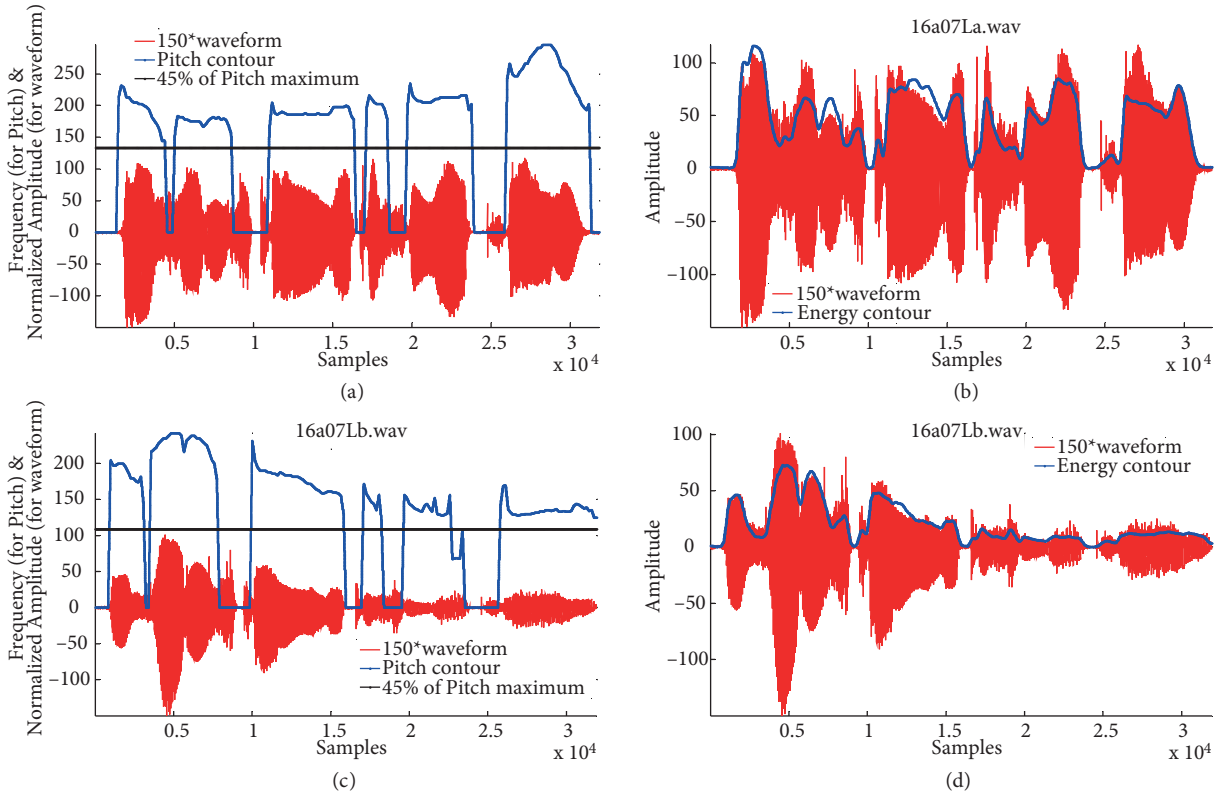

Figure 7. Pitch and energy contour figures of the fifth statement of the EmoDB when it is spoken by the tenth speaker in boredom: a, b) low intensity; c, d) high intensity.

- Feature 58: Mean of the energy duration for the plateaus at minima.
- Feature 64: Maximum of energy duration for the plateaus at maxima.
- Feature 71: Interquartile range of the energy value within the plateaus at maxima.
- Features 201 and 204: Maximum values of the 8th and 11th LPCs, respectively.

As mentioned in the previous section, some of the features reported in Table 1 are very sensitive to the changes of emotional intensities. In Section 4, using different filters and wrappers, appropriate features for application in the 2 frameworks of this manuscript, which are illustrated in Figure 1, are selected.

4. Selecting the best features and classifying utterances for ESR

As expressed in Table 1, 284 features are proposed to be extracted from speech signals. Primary evaluations of the extracted features show that those features indexed by 181, 194, 207, and 220, which were explained before, have zero VC. According to Eq. (12), VC expresses the ratio between variance and mean values.

$$VC = \frac{\sigma}{\mu} * 100\% \tag{12}$$

When VC becomes 0 for a feature, it means that the feature has no information about the emotional contents (or other parameters) of speech signals. Therefore, we delete the aforementioned features from our useful feature set.

Another assessment of the extracted features shows that some of the features related to the variance-related statistical parameters, e.g., 13, 14, 15, and 16, are in the range of 10^5 while some of the rest, for example spectral-related features, are in the order of 10^{-1}. In this way, in order to preserve our classifiers from being misled by large values of features, a linear transformation is applied according to Eq. (13) on the features (ξ), which could fix feature values between 0 and 1.

$$\xi^{\alpha,\beta}_{Normalized} = \frac{\xi^{\alpha,\beta} - Min\left(\xi^{\beta}\right)}{Max\left(\xi^{\beta}\right) - Min\left(\xi^{\beta}\right)}, \begin{cases} \alpha = 1, \ldots, 535 \\ \beta = 1, \ldots, 280 \end{cases} \tag{13}$$

In this equation, α denotes the index of the studied speech signal in the EmoDB dataset while β shows the feature's index in the feature set.

Afterwards, the best features for the classification of speech signals in accordance with their emotional contents should be selected. As mentioned before, VC, as a filter method, is employed to eliminate inappropriate features. In the next step, in order to find suitable features, a wrapper method, e.g., SFFS, is applied. This process has 2 sections: a forward manner, which is correlated with the inclusion of new features to the selected ones, and a backward procedure, which deletes incompatible features [17]. As reported in Table 6, using the SFFS method, the best features for being used in the process of emotional speech classification (ESC) are selected.

Table 6. Selected features for being used for the ESC of EmoDB speech signals to 4 emotional classes, which are anger, happiness, fear, and boredom.

Best features for ESC to 4 classes: angry, happy (joy), fear, and boredom
16, 32, 42, 50, 58, 59, 62, 69, 73, 100, 135, 166, 175, 179, 183, 192, 193, 197, 209, 222, 227, 231, 234, 237–239, 244, 273

Table 7. Confusion matrix resulting from Bayes classifier when 60% of anger, happiness, fear, and boredom-related utterances of the EmoDB are applied for training.

Input speech signal	Classification responses (%)			
	Anger	Happiness	Fear	Boredom
Anger	88.53	4.48	6.99	0
Happiness	9.42	79.21	6.52	4.85
Fear	4.85	3.43	84.40	7.32
Boredom	4.48	2.32	10.41	82.79
Average CCR	83.73			

As illustrated in Figure 1a, after the selection of the best features for being employed for the ESR process, input utterances are classified in accordance with their emotional contents. In this way, Bayes and SVM classifiers are applied in multiclass and binary classification manners, respectively.

Tables 7 and 8 express the consequences of using aforementioned classifiers using the selected features reported in Table 6. CCRs expressed in these tables are attained after using a mathematical averaging on the CCRs that resulted from a 10-fold cross-validation procedure. Figure 8 illustrates the uncertainties observed in these classifications.

5. Classification of the utterances in accordance with their emotional intensities

In the previous section, using the best selected features proposed by the SFFS method and by means of Bayes and SVM classifiers in multiclass and binary classifications, respectively, a framework was applied to classify input utterances with anger, happiness, fear, or boredom emotions to their appropriate emotional classes.

Table 8. CCRs resulting from applying SVM classifier on the 12 possible pairs of emotional classes using selected features when 40% of EmoDB utterances are utilized for testing.

First class	Second class				
	Anger	Happiness	Fear	Boredom	vs. the rest
Anger	-	90.54	95.20	97.28	93.65
Happiness	90.54	-	88.41	97.77	91.10
Fear	95.20	88.41	-	89.39	90.95
Boredom	97.28	97.77	89.39	-	91.73
Average CCR	93.10				91.86

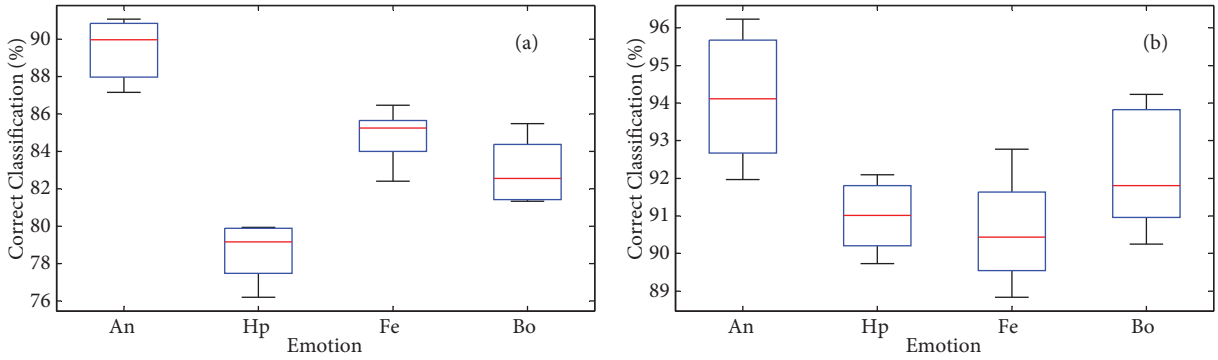

Figure 8. Uncertainties resulting from 10-fold cross-validation processes implemented on the EmoDB utterances when: a) a Bayes classifier is applied to classify anger, happiness, fear, and boredom related utterances in a multiclass classification procedure; (b) a SVM classifier is used to categorize the aforementioned emotional speech signals in a binary manner such that each of these classes is classified against the set of other related utterances.

After this, as displayed in Figure 1, each of the input utterances is classified in conformity with their emotional intensities to high and low levels. As mentioned before, the best features that can be employed to categorize each of the emotional speech signals in accordance with their intensities are mentioned in Table 9.

Table 9. Appropriate features for applying the process of emotional intensity classification. In this table, those features that are appropriate in all of the 4 emotional classes, in (anger and happiness) classes, and in (fear and boredom) classes are illustrated with bold, underlined, and *italic* notations, respectively.

Emotion	Appropriate features for EIC
Anger	25, **26**, **29**, 34, 41, **48**, **64**, 68, **71**, 86, 87
Joy	18, 24, 25, **26**, **29**, 34, 41, **48**, 62, **64**, 65, 68, **71**, 86, 165, 178
Fear	*9*, 18, *23*, **26**, **29**, 39, **48**, 57, *58*, **64**, **71**, *204*
Boredom	*9*, *23*, **26, 29**, **48**, *58*, **64**, **71**, 201, *204*

According to the emotional labels assigned to the input speech signals, appropriate features in agreement with Table 9 are extracted from the utterances. In this way, Bayes, linear, and Gaussian radial basis function (GRBF)-SVM classifiers are applied. In order to improve the fidelity of the results, a 10-fold cross-validation algorithm is used and the average of the obtained CCRs is expressed as the output of that classifier. Additionally, 60% of the data is employed for training and the rest for testing.

Most of the classifiers categorize input utterances with extracted feature vectors (f_1, f_2,..., f_n) to a class indexed by Ξ when their probability model $p(\Xi|f_1, f_2, ..., f_n)$is maximized. Using Bayes' theorem, this model

can be written as in Eq. (14).

$$p\left(\Xi|f_1,f_2,...,f_n\right)=\frac{p\left(\Xi\right)p\left(f_1,f_2,...,f_n|\Xi\right)}{p\left(f_1,f_2,...,f_n\right)} \tag{14}$$

By means of the maximum a posteriori decision rule, extracted feature vectors of (f_1, f_2,..., f_n) are connected to class Ξ in agreement with Eq. (15).

$$\Xi\left(f_1,f_2,...,f_n\right)=\arg\max_{\xi}\left(p\left(\Xi=\xi\right)\prod_{i=1}^{n}p\left(f_i|\xi\right)\right) \tag{15}$$

The results of using the Bayes classifier to categorize input speech signals to high/low emotional intensities are illustrated in Table 10. Afterwards, a linear SVM is applied. In this way, a set of training data (Δ) can be written as in Eq. (16).

$$\Delta=\left\{\left(\mathbf{x}_i,y_i\right)|\mathbf{x}_i\in\mathbf{R}^m,y_i\in\{-1,1\}\right\}_{i=1}^{n} \tag{16}$$

Here, $\mathbf{x}_i$ shows the feature vector extracted from the ith input utterance.

Table 10. CCRs obtained from categorizing input utterances using Bayes classifier in accordance with the emotional intensity content.

Input utterances		Classification responses (%)							
		Anger		Happiness		Fear		Boredom	
		High	Low	High	Low	High	Low	High	Low
Anger	High	63.34	36.66	-	-	-	-	-	-
	Low	39.44	60.56	-	-	-	-	-	-
Happiness	High	-	-	66.47	33.53	-	-	-	-
	Low	-	-	38.11	61.89	-	-	-	-
Fear	High	-	-	-	-	61.34	38.66	-	-
	Low	-	-	-	-	37.47	62.53	-	-
Boredom	High	-	-	-	-	-	-	59.33	40.67
	Low	-	-	-	-	-	-	38.35	61.65
Average CCR(%)		62.14							

In this paper, we assign (+1) to speech signals with high intensity emotions and (–1) for low ones. The consequences of using this classifier are reported in Table 11. As stated in Eq. (17), a GRBF-SVM classifier is utilized to classify emotional speech signals into appropriate intensity levels. Classification results can be found in Table 12.

$$\kappa\left(x_i,x_j\right)=e^{-250\|x_i-x_j\|^2} \tag{17}$$

6. Conclusion

In this paper, we proposed appropriate features for use in the classification of emotional speech signals in accordance with their intensity levels for the first time. Because of the lack of a professional dataset in this field, we focused on finding available datasets with suitable properties that could be applied for this task. We proposed that those datasets that contain utterances related to a sentence spoken with 1 emotional tune by 1 speaker in different intensities could be suitable in this research. Among the available datasets, the EmoDB has the most compatibility with the desired qualifications. In this database, 10 common German sentences are

Table 11. Results of classifying input speech signals using a linear SVM classifier into high and low intensities.

Input utterances		Classification responses (%)							
		Anger		Happiness		Fear		Boredom	
		High	Low	High	Low	High	Low	High	Low
Anger	High	68.45	31.55	-	-	-	-	-	-
	Low	36.72	63.28	-	-	-	-	-	-
Happiness	High	-	-	68.15	31.85	-	-	-	-
	Low	-	-	36.15	63.85	-	-	-	-
Fear	High	-	-	-	-	62.66	37.34	-	-
	Low	-	-	-	-	37.67	62.33	-	-
Boredom	High	-	-	-	-	-	-	65.24	34.76
	Low	-	-	-	-	-	-	35.62	64.38
Average CCR(%)		64.79							

Table 12. CCRs resulting from classification of input utterances using RBF-SVM classifier in accordance with the emotional intensity content.

Input utterances		Classification responses (%)							
		Anger		Happiness		Fear		Boredom	
		High	Low	High	Low	High	Low	High	Low
Anger	High	74.35	25.65	-	-	-	-	-	-
	Low	28.72	71.28	-	-	-	-	-	-
Happiness	High	-	-	78.45	21.55	-	-	-	-
	Low	-	-	26.26	73.74	-	-	-	-
Fear	High	-	-	-	-	70.21	29.79	-	-
	Low	-	-	-	-	31.67	68.33	-	-
Boredom	High	-	-	-	-	-	-	69.84	30.16
	Low	-	-	-	-	-	-	32.52	67.48
Average CCR(%)		71.71							

used. Each of them is spoken by 10 actors including 5 men and 5 women in 7 emotions, which are anger (An), neutral (Nu), fear (Fe), boredom (Bo), happiness (Hp), sadness (Sd), and disgust (Di) [16]. This paper just utilizes An, Hp, Fe, and Bo utterances. In this way, those utterances of the EmoDB that have equal textual content and are expressed by 1 speaker in 1 emotional mood but with different intensity levels (high/low) were studied.

As for the first step, we extracted 284 features from speech signals. After normalizing the values of these features between 0 and 1, features indexed by 16, 32, 42, 50, 58, 59, 62, 69, 73, 100, 135, 166, 175, 179, 183, 192, 193, 197, 209, 222, 227, 231, 234, 237-239, 244, and 273 were selected as the best features for employing for the process of ESR, by using VC and SFFS methods. In this procedure Bayes and SVM classifiers were utilized for multiclass and binary classification of anger, happiness, fear, and boredom emotions, respectively. By means of a 10-fold cross-validation method, these classifiers categorized input utterances to correct emotional classes with CCRs equal to 83.73% and 93.10%, respectively.

Afterwards, in the second framework, which is the main novelty of this paper, the best features for being applied for the categorization of speech signals with 1 emotional type and different intensity levels were studied. These features are reported in Table 9. In order to evaluate the capabilities of these features in the process of EIC, Bayes, linear, and GRBF-SVM classifiers were applied to classify input utterances into the appropriate intensity level (high or low) using these features. In order to improve the fidelity of the results, CCRs were

reported after implementing a mathematical averaging on the outcomes of a 10-fold cross-validation procedure. Thus, the average CCRs that resulted from Bayes, linear, and GRBF-SVM classifiers were reported to be 62.14%, 64.79%, and 71.71%, respectively.

References

[1] Busso C, Lee S, Narayanan S. S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE T Audio Speech 2009; 17: 582-596.

[2] Yun S, Yoo CH. Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In: ICASSP 2009 Proceedings; 19–24 April 2009; Taipei, Taiwan. New York, NY, USA: IEEE. pp. 4169-4172.

[3] Bezooijen RV. The Characteristics and Recognizability of Vocal Expression of Emotions. Dordrecht, the Netherlands: Foris Publications, 1984.

[4] Tolkmitt FJ, Scherer KR. Effect of experimentally induced stress on vocal parameters. J Exp Psychol Human 1986; 12: 302-313.

[5] McGilloway S, Cowie R, Douglas-Cowi E. Approaching automatic recognition of emotion from voice: a rough benchmark. In: ISCAWSE 2000 Proceedings; 5–7 September 2000; Newcastle, UK.

[6] Hammal Z, Bozkurt B, Couvreur L, Unay U, Caplier A, Dutoit T. Passive versus active: vocal classification system. In: EUSIPCO 2005 Proceedings; 4–8 September 2005; Antalya, Turkey. New York, NY, USA: IEEE. pp. 1-4.

[7] Ververidis D, Kotropoulos C, Pitas I. Automatic emotional speech classification. In: ICASSP 2004 Proceedings; 17–21 May 2004; Montreal, Canada. New York, NY, USA: IEEE. pp. 593-596.

[8] Pao TL, Liao WY, Chien CS, Chen YT, Yeh JH, Cheng YM. Comparison of several classifiers for emotion recognition from noisy Mandarin speech. In: IIH-MSP 2007 Proceedings; 26–26 November 2007; Kaohsiung, Taiwan. New York, NY, USA: IEEE. pp. 23-26.

[9] Yang C, Pu X. Efficient speech emotion recognition based on multisurface proximal support vector machine. In: RAM 2008 Proceedings; 21–24 September 2008; Chengdu, China. New York, NY, USA: IEEE. pp. 55-60.

[10] Hansen JHL, Wooil K, Rahurkar M, Ruzanski E, Meyerhoff J. Robust emotional stressed speech detection using weighted frequency subbands. Eurasip J Adv Sig Pr 2011; 2011: 906789.

[11] Tawari A, Trivedi M. Speech emotion analysis in noisy real-world environment. In: ICPR 2010 Proceedings; 23–26 August 2010; İstanbul, Turkey. New York, NY, USA: IEEE. pp. 4605-4608.

[12] Kim W, Hansen JHL. Angry emotion detection from real-life conversational speech by leveraging content structure. In: ICASSP 2010; 14–19 March 2010; Dallas, TX, USA. New York, NY, USA: IEEE. pp. 5166-5169.

[13] Karimi S, Sedaaghi MH. Robust emotional speech classification in the presence of babble noise. International Journal of Speech Technology 2013; 16: 215-227.

[14] Song M, Chen C, Bu J, You M. Speech emotion recognition and intensity estimation. Lect Notes Comp Sci 2004; 3046: 406-413.

[15] Banziger T, Tran V, Scherer KR. The Geneva emotion wheel: a tool for the verbal report of emotional reactions [poster]. In: ISRE 2005 Proceedings; Bari, Italy, 2005.

[16] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of German emotional speech. In: Interspeech 2005 Proceedings; Lisbon, Portugal, 2005.

[17] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun 2006; 48: 1162-1181.

[18] Sedaaghi MH. Gender classification in emotional speech. In: Mihelic F, Zibert J, editors. Speech Recognition: Technologies and Applications. Rijeka, Croatia: InTech, 2008. pp. 363-376.

[19] Chang CC, Hsu CW, Lin CJ. The analysis of decomposition methods for SVM. IEEE T Neural Networ 2000; 11: 1003-1008.

[20] Ratsch G, Mika S, Scholkopf B, Muller KR. Constructing boosting algorithms from SVMs: an application to 1 class classification. IEEE T Pattern Anal 2002; 24: 1184-1199.

[21] Chang JH, Kim NS, Mitra SK. Voice activity detection based on multiple statistical models. IEEE T Signal Proces 2006; 54: 1965-1976.

[22] Rabiner LR, Sambur MR. Voiced-unvoiced-silence detection using Itakura LPC distance measure. In: ASSP 1977 Proceedings; May 1977. New York, NY, USA: IEEE. pp. 323-326.

[23] Wechsler JD. Detection of human speech in structured noise. In: ASSP 1994 Proceedings, 19–22 April 1994; Adelaide, Australia. New York, NY, USA: IEEE. pp. 237-240.

[24] Beritelli F, Casale S, Cavallaro A. A robust voice activity detector for wireless communications using soft computing. IEEE J Sel Area Comm 1998; 16: 1818-1829.

[25] Snell RC, Milinazzo F. Formant location from LPC analysis data. IEEE T Speech Audi P 1993; 1: 129-134.

[26] Markel JD, Gray AH. Linear Prediction of Speech. Berlin, Germany: Springer-Verlag, 1976.

[27] Rabiner LR, Schafer RW. Digital Processing of Speech Signals. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[28] Loizou P. COLEA: A MATLAB Software Tool for Speech Analysis. Fayetteville, AR, USA: University of Arkansas, 2003.

[29] Hess WJ. Pitch and voicing determination. In: Furui S, Sondhi MM, editors. Advances in Speech Signal Processing. New York, NY, USA: Marcel Dekker, 1992. pp. 3-48.

[30] Sondhi MM. New methods of pitch extraction. IEEE T Acoust Speech 1968; 16: 262-266.

[31] Ververidis D, Kotropoulos C. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In: EUSIPCO 2006 Proceedings; 4–8 September 2006; Florence, Italy. New York, NY, USA: IEEE. pp. 1-5.

[32] Shah F, Krishnan V, Sukumar R, Jayakumar A, Anto B. Speaker independent automatic emotion recognition from speech, a comparison of MFCCs and discrete wavelet transforms. In: ARTCC 2009 Proceedings; 27–28 October 2009; Kottayam, India. New York, NY, USA: IEEE. pp. 528-531.

[33] Ning T, Whiting S. Power spectrum estimation via orthogonal transformation. In: ASSP 1990 Proceedings. New York, NY, USA: IEEE. pp. 2523-2526.

[34] Hermansky H. Perceptual linear predictive (PLP) analysis for speech. J Acoust Soc Am 1990; 1: 1738-1752.

[35] Hermansky H, Morgan N, Bayya A, Kohn P. RASTA-PLP speech analysis technique. In: ICASP 1992; 23–26 March 1992; San Francisco, CA, USA. New York, NY, USA: IEEE. pp. 121-124.