EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Efficiency of chosen speech descriptors in relation to emotion recognition

Dorota Kamińska[1*], Tomasz Sapiński[1] and Gholamreza Anbarjafari[2,3]

Abstract

This research paper presents parametrization of emotional speech using a pool of common features utilized in emotion recognition such as fundamental frequency, formants, energy, *MFCC*, *PLP*, and *LPC* coefficients. The pool is additionally expanded by perceptual coefficients such as *BFCC*, *HFCC*, *RPLP*, and *RASTA PLP*, which are used in speech recognition, but not applied in emotion detection. The main contribution of this work is the comparison of the accuracy performance of emotion detection for each feature type based on the results provided by both *k*-NN and SVM algorithms with 10-fold cross-validation. Analysis was performed on two different Polish emotional speech databases: voice performances by professional actors in comparison with the author's spontaneous speech.

Keywords: Voice, Emotion recognition, Perceptual coefficients, Speech signal analysis

1 Introduction

Emotion recognition methods utilize various input types, i.e., facial expressions [1], speech, gestures, and body language [2], and physical signals such as electrocardiogram (ECG), electromyography (EMG), electrodermal activity, skin temperature, galvanic resistance, blood volume pulse (BVP), and respiration [3]. Speech is most accessible from the aforementioned signals. Therefore, much research in the field of emotion recognition is focused on human voice.

According to Plutchik's theory [4], there are eight primary bipolar emotions: joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation. These emotions are biologically primitive and have evolved in order to increase the reproductive fitness. Primary emotions can be expressed at different intensities and can be mixed with one another to form different emotional states. This translates to the perception of natural emotions, which is a complex and subjective process. Recognizing several different emotional states in a given situation is very common.

Initially, research on emotion recognition was mostly conducted using acted-out speech which carried undisturbed and clear singular emotion expressions [5]. In 2009, at the Affective Computing and Intelligent Interaction (ACII) conference, a session was held dedicated to emotion recognition from ambiguous samples (containing a mixture of emotions). This started a new wave in the field of emotion recognition in which the researchers abandoned acted-out speech in favor of spontaneous speech [6]. In line with that, this article describes a database of Polish emotional speech extracted from natural discussions in TV programs. The database consists of over 784 samples divided into seven sets representing primary emotional states, although based on Plutchik's wheel, which presents eight basic emotions, the psychologists who were involved in labeling the data did not label any of the audio signals as "trust." Hence, in this work, only seven basic emotions, namely joy, fear, surprise, sadness, disgust, anger, and anticipation, are used. Moreover, for comparative purpose, emotions performed by professional actors were analyzed.

Emotion recognition from speech is a pattern recognition problem. Therefore, standard pattern recognition methodologies, which involve feature extraction and classification, are used to solve the task [7]. The number of speech descriptors that are being taken into consideration is still increasing. Mostly acoustic and prosodic features from the set of Interspeech 2009 Challenge [8] are utilized. Therefore, fundamental frequency, formants, energy, Mel

*Correspondence: dorota.kaminska@p.lodz.pl
[1] Institute of Mechatronics and Information Systems, Stefanowskiego 18/22, 90-924 Lodz, Poland
Full list of author information is available at the end of the article

Frequency Cepstral Coefficients (MFCC), or Linear Prediction Coefficients (LPC) are widely explored. Nevertheless, the search for new speech features is ongoing.

This research is conducted using a pool of commonly used features utilized in emotion recognition, such as fundamental frequency f_0, formants, energy, *MFCC*, Perceptual Linear Prediction (*PLP*), and *LPC* coefficients. The pool is additionally expanded by perceptual coefficients, such as Bark Frequency Cepstral Coefficients (*BFCC*), Human Factor Cepstral Coefficients (*HFCC*), Revised Perceptual Linear Prediction (*RPLP*), and RASTA Perceptual Linear Prediction (*RASTA PLP*). The main contribution of this work is to test abovementioned perceptual features, which are applied in speech recognition research but omitted in emotion recognition. All feature sets were tested separately in order to demonstrate their impact on emotion recognition. The verification of feature set efficiency was carried out using both *k-NN* and Multi Class Support Vector Machine (*SVM*) [9] with radial kernel classifiers applying 10-fold cross-validation on two independent speech corpora.

The outline of this paper is as follows. Section 2 describes the emotional speech database: structure of the corpora, methods for selecting of the recording source, and the process of emotional speech labeling. Section 3 introduces the examined speech descriptors. Section 4 presents obtained results. Finally, Sections 5 and 6 concludes and summarizes the paper.

2 Emotional speech corpora

Emotional speech samples can be divided into three categories, taking into account their source: spontaneous, invoked, and acted or simulated emotions. The first type can be obtained by recording speakers in natural situations or using TV programs such as talk shows, reality shows, or various types of live coverage. This type of material might not always be of satisfactory quality (background noise, artifacts, overlapping voices, etc.) and may obscure the exact nature of recorded emotions. Moreover, collections of spontaneous speech must be evaluated by human decision makers to determine the gathered emotional states.

Another method of sample acquisition is provoking an emotional reaction by using drugs or staged situations. Appropriate states are induced using imaging methods (videos, images), stories, or computer games. This type of recording is preferred by psychologists, although the method cannot provide desirable effects as reaction to the same stimuli may differ. Similar to spontaneous speech recordings, triggered emotional samples should be subjected to a process of identification by independent listeners.

The third source of emotional speech is acted-out samples. Speakers can be both actors as well as unqualified volunteers. This type of material is usually comprised of high-quality recordings, with clear undistorted emotion expression. Furthermore, the ease of acquiring recordings opens a possibility of obtaining several utterances, representing different emotional states from a single user. However, the acoustic characteristics of such an utterance may be exaggerated, while more subtle features are completely ignored.

2.1 Polish spontaneous speech database

Based on Robert Plutchik's theory, a corpus not only consists of primary emotions with the addition of complex emotional states but also consists of a much wider range than the commonly used databases [10]. The first step was to gather audio samples containing the emotional carrier of basic states from Plutchik's wheel of emotions: joy, sadness, anger, fear, disgust, surprise, and anticipation. All samples were assessed by a large group of human evaluators (experts and volunteers) and labeled into the abovementioned classes of emotions, summarized in Table 1.

Statistical analysis is an integral part of creating an emotional speech corpus. It should meet certain criteria. One of them is to preserve the distribution of the parameters (characteristics) of the subject of research relevant to the application, which affects its reliability. The set, presented in this study, is a collection of emotional expressions in Polish. Spatial extent, time, place, and personal characteristics of the speaker are not restricted.

The selection of a representative sample recordings is one of the key elements affecting the research credibility. It is assumed that a sample is representative when all the values which could affect the test results are present. Because the process of emotion expression is subjective, depending primarily on gender and age, these variables are taken into account in the process of the corpora creation. In order to retain the right proportions of these variables (almost equal), the abovementioned information is one of the guidelines used when selecting speech sources. This assumption is largely limited due to a lack of personal data of the speakers in recordings obtained from radio auditions.

Table 1 Structure of the corpus

Emotion name	Type	Number	Gender
Anger	Primary	111	55 female/56 male
Anticipation	Primary	88	44 female/44 male
Joy	Primary	165	90 female/33 male
Fear	Primary	48	26 female/22 male
Surprise	Primary	128	61 female/67 male
Sadness	Primary	115	57 female/58 male
Disgust	Primary	90	48 female/42 male

The most important feature of the samples is the authenticity of presented emotional states, which narrows the search area. The authors focus mainly on the materials from live shows and programs such as *reality show*. The reactions and feelings presented by the participants of such programs seem spontaneous and provoked by events and discussions. For example, shows presenting political and social problems (e.g., *Państwo w Państwie* by Polsat TV) contain a large number of anger displays. The assumption of authenticity of emotions could be false and is associated with the subjective evaluation performed by the authors and volunteers involved in the labeling process. What is important to mention is the fact that the collected recordings often contain background noise, which also might have affected the assessment.

The emotional state of the speaker can be identified based on short utterances such as *Yes* or *No* [11]. Thus, short sentences, or even single words, are suitable for emotional analysis. Occasionally, additional sounds such as screaming, squealing, laughing, or crying carry the information about the speaker's emotional state. Therefore, in addition to full spoken words, such sounds which occur in everyday communication are featured in the created corpora.

In addition, a neutral speech model (without emotional coloring) is created for the purposes of emotional research. It is composed of statements from [12] and supplemented with statements of journalists commenting on various events. Such utterances are usually neutral and do not carry any emotional load. This model consists of 235 statements and is not subjected to labeling by volunteers.

The labeling process is divided into two parts. First, the recordings are divided into seven groups (basic emotions). This process is conducted by the authors and students of the faculty of psychology from the University of Lodz. The division is performed with the use of video material which allows access not only to voice and semantics but also to the visual display of emotions, such as gestures or facial expressions. In the second part of the process, the volunteers label the samples based on audio input only. This emphasizes how subjective the perception of emotions really is.

Listening to pre-qualified samples is performed to test whether the listener is able to identify the emotional content of the recording. The volunteer group consists of 15 people, both male and female, aged 21 to 58 years with no hearing disabilities. The task is to assess the recordings and classify them into the groups of seven basic emotions. All listeners are presented with a random set of samples that consists of at least half of each pre-qualified basic emotion recording. The evaluators listen to audio samples one by one, and each assessment is recorded in the database. Every sample could be replayed a number of times before the final decision, but after the classification, it is not possible to return to the recording.

Average recognition amounted to 82.66% in the range of 63 to 93%. However, it should be noted that the prequalified samples rated by the authors and students of psychology are the base of the classification and that assessment is also subjective. Therefore, the samples which repeatedly mismatched the labels of the pre-qualification are incorporated into forming ambiguously defined states. Emotions, assessed identically by at least 10 people, are classified as pure prototype states. The database can be made available upon request, for research purposes only.

2.2 Polish acted speech database

The Polish acted emotional speech is made available by the Medical Electronics Division, Technical University of Lodz. This database consists of 240 sentences uttered by eight speakers (four males and four females). Recordings for every speaker were made during a single session. Each speaker utters five different sentences with six types of emotional load: joy, boredom, fear, anger, sadness, and neutral (no emotion). Recordings were taken in the aula of the Polish National Film Television and Theater School in Lodz.

Methodology of inducing a particular emotional attitude follows recommendations [13]: the uttering of each database sentence (sentences have no particular emotional meaning) is preceded by uttering a sentence with a clear emotional connotation, relevant for the current recording.

To assess a quality of the database material, the recordings are evaluated by 50 subjects, through a procedure of classification of 60 randomly generated samples (10 samples per particular emotion). Listeners are asked to classify each utterance into emotional categories. An average rate of correct recognition for this evaluation experiment is 72% (ranging from 60 to 84% for different subjects) [12].

3 Methods

3.1 Prosodies

F_0 is the frequency of vocal folds. It is responsible for the scale of the human voice and accent. It plays an important role in the intonation, which has a significant impact on the nature of the speech. F_0 changes during articulation. The rate of those changes depends on the speaker's intended intonation [14]. There are many methods to determine the fundamental frequency. In this paper, f_0 is extracted using the autocorrelation method. The analysis window is set to 20 ms with 50% overlap. It is difficult to objectively assess the behavior of f_0 based on the chart. Therefore, statistical parameters related to f_0 are extracted.

Formant frequencies are the frequencies at which the local maxima of the speech signal spectrum envelope occur. They are the properties of the vocal tract. Based on this, it is possible to determine who the speaker is and about what and how he/she is speaking [15]. In practice, applications from three to five formants are used. In this paper, three formant frequencies are estimated. On their basis, parameters such as mean, median, standard deviation, maximum, and minimum are determined. A total of 15 features are extracted.

Speech signal energy, which refers to the volume or intensity of speech, also provides information that can be used to distinguish emotions (i.e., joy and anger increase energy levels in comparison to other emotional states).

3.2 Spectral coefficients

The perceptual approach is based on frequency conversion, corresponding to the subjective reception of the human hearing system. For this purpose, the perceptual scales such as Mel or Bark are used. In this paper, Mel Frequency Cepstral Coefficients (MFCC) [16], Human Factor Cepstral Coefficients (HFCC) [17], Bark Frequency Cepstral Coefficients (BFCC) [18], Perceptual Linear Prediction(PLP) [19], RASTA Perceptual Linear Prediction (RASTA PLP) [20], and Revised Perceptual Linear Prediction (RPLP) [18] coefficients are taken into consideration. The entire scheme for perceptual feature extraction is shown in Fig. 1.

4 Efficiency of features

The verification of the efficiency of feature subsets is carried out using k-NN and SVM classifiers applying *10-fold cross-validation* on two independent speech corpora. This method allows to correctly evaluate descriptor efficiency. It is based on a random division of a whole set into 10 subsets of equal size. Then, a single subset is used as the test set, and the rest acts as the training set. This process is repeated 10 times, so that every subset is used as a test set. The final result is achieved by calculating the average of the results of each iteration. In this way, dominating features are distinguished. Table 2 presents the efficiency of commonly used feature subsets. In the course of research, the value of k was selected to achieve the highest classification results.

4.1 Perceptual coefficients

The next step of analysis includes a detailed comparison of perceptual coefficient efficiency. As in the previous step, the classification is carried out using the k-NN and SVM classifiers applying 10-fold cross-validation. The number of perceptual coefficients giving maximum results depends on the type of examined features; it was selected in order to achieve the highest classification results. The value of k in case of k-NN algorithm is experimentally chosen in order to give the highest classification results for a given group. For each signal frame, proper coefficients are obtained, and basing on those, statistical features are extracted.

Subsequently, the sets of the aforementioned coefficients are expanded by their dynamic parameters. Classification efficiency with a various combination of these parameters is shown in Table 3. In both cases, none of dynamic parameters provide an increase in recognition rate.

It can be noticed that perceptual coefficients provide much higher recognition results than previously tested features. The best results in both corpora are obtained by using hybrid coefficients. In case of k-NN, the highest recognition was achieved for BFCC, 56.5% for acted speech and 74.5% for natural speech. Generally, the accuracy performance for SVM is much lower. However, the case of k-NN proves the validity of hybrid coefficient application. The best results were obtained using RPLP coefficients 43.88% for acted speech and 59.4% for neutral speech using BFCC coefficients.

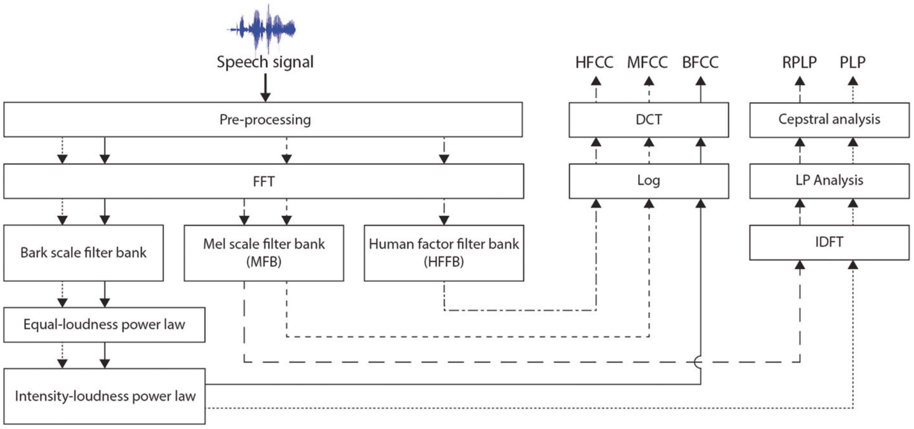

Fig. 1 Feature extraction process for different methods

Table 2 Average recognition results [%] of commonly used features subsets: acted speech (*AS*) and natural speech (*NS*)

Feature set	*k*-NN AS	*k*-NN NS	SVM AS	SVM NS
f_0	31.6	39.8	19.66	26.98
$f_1 - f_3$	38.8	39.5	16.88	27.5
Energy	30.4	47	29.95	46.87
LPC	36.7	57.8	29.91	60.51

4.2 Selection

Selection is conducted on specific subsets to improve the efficiency of classification. For the purpose of this research, two different feature selection techniques were chosen. The first is the Fisher-Markov Selector (FMS) [21], which is independent of the classifier. The second, wrapper method, is a classifier dependent selection—Sequential Forward Selection (SFS) [22]. The results of experiments for both corpora are presented in Table 8.

Table 3 Classification efficiency using *k*-NN and SVM [%] with various combinations of dynamic parameters for acted speech (AS) and natural speech (NS)

Coefficients	AS *k*-NN	AS SVM	NS *k*-NN	NS SVM
MFCC	56.5	40.5	71.7	58.31
MFCC + ΔMFCC	52.7	40.5	68.7	42.91
MFCC + ΔΔMFCC	51.1	37.13	66.9	42.91
MFCC + ΔMFCC + ΔΔMFCC	36.7	33.75	66.9	37.5
HFCC	52.7	39.66	70.6	55.17
HFCC + ΔHFCC	47.7	37.13	67.8	40.32
HFCC + ΔΔHFCC	46	36.7	66.9	36.23
HFCC + ΔHFCC + ΔΔHFCC	43.9	36.7	64.6	34.59
BFCC	56.5	40.92	74.5	59.4
BFCC + ΔBFCC	56.5	36.28	70.25	44
BFCC + ΔΔBFCC	53.6	35.86	73.4	38.41
BFCC + ΔBFCC + ΔΔBFCC	53.6	35.86	69.9	36.28
PLP	54	21.51	71.4	36.83
PLP+ ΔPLP	52.7	16.87	66.9	22.47
PLP + ΔΔPLP	53.6	16.87	66.9	22.47
PLP + ΔPLP + ΔΔPLP	50.2	16.87	64.7	22.47
RPLP	55.7	43.88	70	54.76
RPLP + ΔRPLP	54.4	40.08	69.3	51.9
RPLP + ΔΔRPLP	52.3	40.92	68.4	51.9
RPLP + ΔRPLP + ΔΔRPLP	52.3	40.5	68	51.9
RASTA PLP	43	36.28	52.2	50
RASTA PLP + ΔRASTA PLP	40.1	35.02	51.8	36.83
RASTA PLP + ΔΔRASTA PLP	40.9	32.48	48.5	36.83
RASTA PLP + ΔRASTA PLP + ΔΔRASTA PLP	40.1	32.48	48.1	36.83

Table 4 Accuracy performance of emotion recognition [%] without selection (–) in comparison with two different selection methods (FMS and SFS) and extraction (S-PCA) methods using *k*-NN algorithm as a classifier

	Acted speech				Natural speech			
	–	SFS	FMS	S-PCA	–	SFS	FMS	S-PCA
f_0	31.6	38.4	40.92	32.9	39.8	40.9	46.58	40.9
$f_1 - f_3$	38.8	39.2	38.8	34.2	39.5	39.6	45.73	37.1
Energy	30.4	36.7	36.7	31.6	47	47.7	55.96	44.6
LPC	36.7	36.7	36.7	36.3	57.8	57.8	63.06	48.5
BFCC	55.7	64.4	64.4	51.5	76.7	77.7	76.7	71.1
MFCC	56.5	59.9	58.10	40.5	71.7	74	74	62.7
HFCC	51.1	52.7	52.7	45.6	70.4	72.9	75.61	68.8
PLP	54	54.4	57.25	51.9	71.4	71.4	71.4	70.4
RPLP	55.7	58.6	59.9	51.9	70	71.3	71.38	64.2
RASTA	43	46.8	49.36	35	52.2	52.2	72.76	44.1

4.3 Comparison of emotion recognition quality utilizing feature selection, applied on feature subsets

Tables 4 and 5 present the results of emotion recognition, after applying *SFS* and *FMS* methods on feature subsets. Additionally, both methods are compared with feature extraction *S-PCA* [23]. The results are presented for both corpora with division into subsets of attributes.

Analyzing results obtained with *k*-NN algorithm, in most cases, one can see a recognition rate improvement after using the SFS or FMS selection methods. For SFS, the exceptions are LPC in the case of the acted speech database, and PLP and RASTA PLP in the case of

Table 5 Accuracy performance of emotion recognition [%] without selection (–) in comparison with two different selection methods (FMS and SFS) and extraction (S-PCA) methods using SVM algorithm as a classifier

	Acted speech				Natural speech			
	–	SFS	FMS	S-PCA	–	SFS	FMS	S-PCA
f_0	19.66	19.66	19.66	34.17	26.98	26.98	26.98	52.55
$f_1 - f_3$	27.5	27.5	25.31	40.92	16.877	17.29	27.5	57.38
Energy	29.95	29.95	29.95	35.86	46.87	46.87	47.72	53.97
LPC	29.91	32.49	32.06	32.06	60.51	65.35	60.79	64.20
BFCC	40.92	47.67	46.83	56.11	59.4	62.53	59.40	81.7
MFCC	40.5	45.14	41.35	54.96	58.31	63.48	58.31	77.92
HFCC	39.66	46.83	40.08	57.38	55.17	56.13	55.17	77.79
PLP	22.47	22.47	22.47	46.83	21.51	22.36	22.47	73.56
RPLP	43.88	44.72	44.72	57.80	54.76	59.12	54.76	70.31
RASTA	36.28	39.66	36.28	51.9	50	60.04	62.53	73.84

the spontaneous speech database. The recognition rates achieved for these attributes, after applying SFS, remained unchanged. An improvement of recognition can be also observed after applying FMS, with the exception of $f_1 - f_3$ and LPC in the case of acted speech and BFCC and *PLP* in the case of the spontaneous speech database. The recognition rates achieved for these attributes, after applying FMS, remained unchanged.

After applying feature extraction (S-PCA) for both corpora, in most cases, a slight decrease of recognition rate could be observed in comparison to results obtained without any selection. Moreover, the results never reach values higher than ones obtained using one of the selection methods.

In case of SVM, the recognition rate is improved by using the SFS and FCBS. The exceptions for SFS are f_0, $f_1 - f_3$, energy, and PLP in the case of the acted speech database and, in the case of the spontaneous speech database, f_0 and energy. The recognition for these attributes remained unchanged after selection. For FMS, the exceptions are f_0, energy, PLP, and *RASTA PLP* in the case of the acted speech database and, in the case of the spontaneous speech database, f_0, BFCC, MFCC, HFCC, and RPLP. The recognition for these attributes remained unchanged after selection. Using FMS resulted in slight decrease of the recognition rate in case of $f_1 - f_3$ in the case of the acted speech database.

In contrast to *k*-NN, after applying feature extraction for both corpora, in most cases, a huge increase of recognition rate could be observed in comparison to results obtained with and without both selection methods. Significantly higher results can be seen for spontaneous speech database.

The best results are achieved for the subset containing BFCC coefficients (81.7% for with S-PCA). The lowest results are obtained in the case of formants and PLP coefficients: 17.29% for SVM with SFS and 22.36% for SVM with SFS, respectively.

In the case of acted speech corpora, the highest results are achieved for BFCC: 64.4% (*k*-NN with SFS and FMS), and the lowest results are achieved by using fundamental frequency and PLP coefficients: 19.66% (SVM with SFS/FMS) and 22.47% (SVM with SFS/FMS), respectively.

4.4 Comparison of emotion recognition quality utilizing feature selection, applied on combined feature set

Tables 6 and 7 present results of emotion recognition conducted on a feature set comprised of all subsets presented in Table 4 or 5, for acted and natural speech accordingly. All selection and extraction methods were applied exactly as in the previous section. Additionally, both feature selection and extraction are applied on the combined feature set (SFS + S-PCA, FMS + S-PCA).

Table 6 Accuracy performance of emotion recognition [%] for acted speech

Clasifier	–	SFS	FMS	S-PCA	SFS + S-PCA	FMS + S-PCA
k-NN	60.8	64.55	65.8	34.17	53.59	56.11
SVM	16.87	16.87	42.19	51.47	54.43	56.54

– without selection, *FMS and SFS* two different selection methods, *S-PCA* extraction, *SFS + S-PCA and FMS + S-PCA* selection and extraction

For the acted database, the best results for acted speech were obtained for *k*-NN with FMS (65.8%). The lowest accuracy rate was achieved for SVM, 16.87% for the whole feature set and after applying SFS.

The highest accuracy performance for natural speech was achieved with SVM. After applying both selection and extraction (FMS + S-PCA), it reached 83.95%. Similar to the previous database, the lowest results were obtained for SVM with SFS and for the whole feature set.

Applying selection reduced the feature set size from initial 448 attributes to 57 (for SFS) and 99 (for FMS), for the acted speech corpora. For the spontaneous database, the values decreased to 88 (SFS) and 90 (FMS) from the initial 473. Distribution of each feature subset, after applying selection to the whole feature set, is presented in Fig. 2.

5 Discussion

Expression of emotion generally depends on the speaker, the culture, and environment [24]. In order to reduce those factors, both emotional speech databases used in this research contain utterances in Polish, performed by native Polish speakers.

One can notice significantly lower recognition results for the acted speech database. This is the result of the different contents of the two databases. In case of acted speech corpora, the relatively low number of speakers, in comparison to the sample count, could affect performance of the classifiers. Moreover, the contents of the utterances were the same throughout different emotional states; this is an advantage if one wants to ensure that human judgment on the perceived emotion is solely based on the emotional content [25]; however, in case of an automated recognition system, the line between emotion and speech recognition might be blurred, specially in this case, where the tested features are commonly used in speech recognition tasks.

Table 7 Accuracy performance of emotion recognition [%] for natural speech

Clasifier	–	SFS	FMS	S-PCA	SFS + S-PCA	FMS + S-PCA
k-NN	78.9	78.9	80.25	51.30	68.72	77.92
SVM	22.49	22.49	70.91	72.83	78.87	83.95

– without selection, *FMS and SFS* two different selection methods, *S-PCA* extraction, *SFS + S-PCA and FMS + S-PCA* selection and extraction

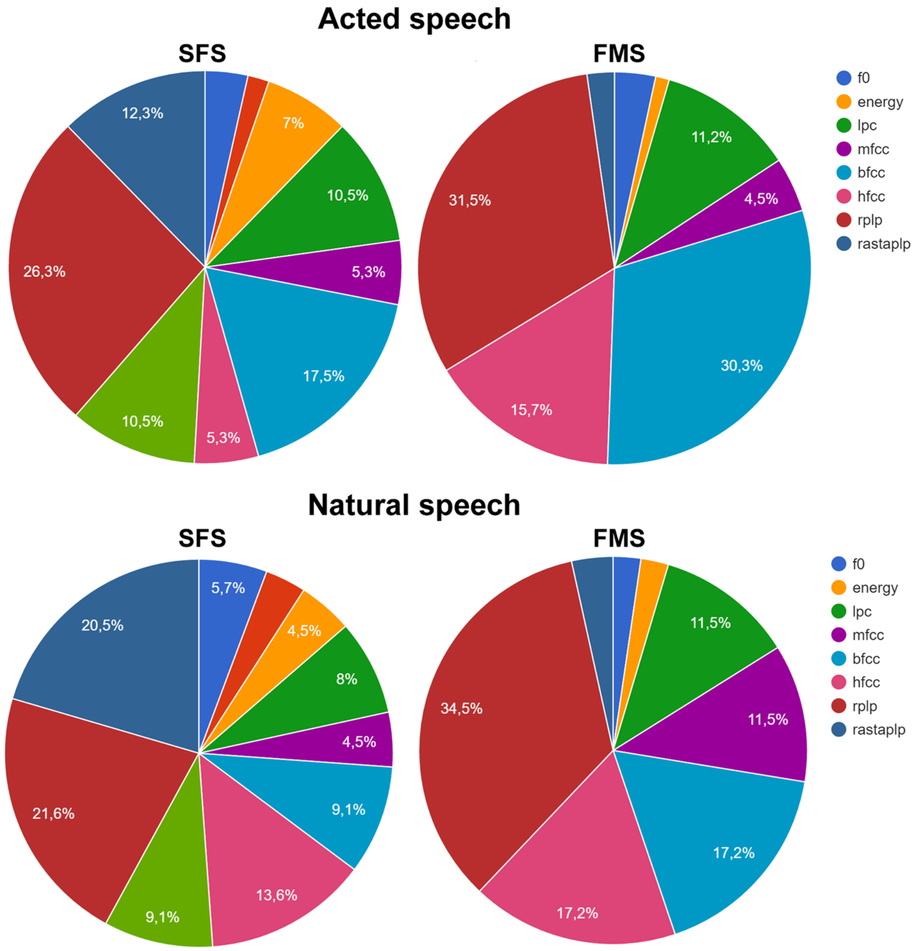

Fig. 2 Distribution of feature groups after applying selection on the combined feature set (*up* acted speech, *down* natural speech, *left* SFS, *right* FMS)

The natural speech corpus contents were selected in order to ensure a proper number and variety of samples which considerably increased the recognition rates. The high number of speakers, as well as different contents of utterances, guaranteed that the extracted features carried information about emotion and not just about the speech or the speaker. Moreover, the higher number of samples in the natural speech database had a reflection in classification results.

For both corpora, similar discrepancies were observed for classification performed by human deciders. Average human recognition rate for natural and acted speech amounted to 81 and 72%, respectively.

The most important issue in emotion speech recognition is the extraction of discriminative features that efficiently characterize different emotional state. It is believed that a proper selection of features significantly affects the classification performance.

After applying SFS and FMS on the combined feature set for both databases, presented in Fig. 2, one can notice that the majority of selected features come from the RPLP, BFCC, and HFCC subsets, with a supplement from RASTA PLP, LPC, and MFCC groups. This correlates with the recognition results for each individual group of features, presented in Tables 4 and 5. Accuracy performance achieved in this research justifies adopting those features for the purpose of emotion recognition.

What is also worth mentioning is the fact that none of the dynamic features, presented in Table 3, were included in the feature sets obtained after applying both selection methods on combined feature groups. They were also excluded after executing selection on isolated feature subsets, as shown in Table 8. The addition of dynamic features, even to feature subsets, did not improve classification results; in fact, in most cases, it reduced the accuracy. This behavior might be caused by a great increase of feature space, where the dynamic features are either nondescriptive or redundant and act more as noise than a carrier of emotional information.

Selecting an appropriate feature reduction method is a crucial step in the recognition process; however, feature selection and classification is highly dependant on the

Table 8 Features after subset selection SFS and FMS for both corpora: acted *AS* and natural *NS* speech

	AS *k*-NN SFS	AS SVM SFS	AS FMS	NS *k*-NN SFS	NS SVM SFS	NS FMS
f_0	Mean; quartile: 1, 3; median; interquartile; skewness; variation rate	Mean; median; std; range; interquartile; kurtosis; variation rate; rising and falling slope mean	Mean; max; min; quartile: 1, 3; interquartile; kurtosis; median; range; rising range max/mean; falling range max/min; std; rising slope max/min;	Mean; median; max; range; quartile: 1, 3; interquartile; kurtosis; skewness; variation rate	Mean; median; range	Max; mean; std; median, range; quartile: 1, 3; interquartile; rising range max/mean; falling range mean
$f_1 - f_3$	Mean: 1, 2; max: 1, 3; median: 2; min: 2, 3; std: 3	Mean: 1; max: 1	Mean: 2, 3; max: 2, 3; min: 2, 3; median: 2, 3	Mean: 1, 2; max: 2; median: 1, 2; min: 3; std: 3	Mean: 1; max: 1	Mean: 3; max: 1, 2, 3; median: 2, 3; min: 2, 3; std: 1
En	Mean, median, std	Max, min, range	Median, std	Max, min, median, std, range	Max, min, median, std, range	Max, mean, std, range
LPC	Mean: 2–11	Mean: 5, 6, 7, 12	Mean/median /max/min: 3–5;	Mean: 2, 3, 4, 7, 9; median: 2	Mean: 2, 4, 5, 7; median: 7	Mean/median /max/min: 3–5;
BFCC	Mean: 1, 6, 10; median: 1, 6; max: 1; min: 1, 6	Median: 1, 3; max: 1, 3, 4; min: 1, 6, 8	Mean: 1, 3, 5, 6; median: 1, 5, 6, 7; min: 1, 2, 4, 3, 5, 6, 7, 8	Mean: 1, 4, 5, 6, 7, 8, 9, 11, 12; median: 1, 2, 3, 4, 5, 8; std: 1; max: 1, 4, 12; min: 2	Mean: 7; median: 2, 3, 5; std: 1, 2, 3, 4, 5; max: 1, 4, 5, 6, 10, 12; min: 1, 3, 4, 5, 6, 10	Mean: 1, 4, 5, 6; median: 1, 2, 4, 3, 5, 7; max: 1, 2, 3, 4, 5, 6, 7, 8; min: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
MFCC	Mean: 1, 6, 10, 11, 12; std: 12; median: 1, 6; max: 1, 7; min:1	Mean:1; max: 1, 11, 12; min: 8, 10, 12	Mean: 1; std: 1; median: 1; max: 1, 4, 11, 12, 14; min: 1, 5, 6, 7, 8, 9, 10, 11	Mean: 1, 2, 3, 4, 6, 9, 10; std: 1, 7, 11, 12; median: 3, 4, 7; max: 1, 2, 4, 5; min: 1, 3, 5, 8, 12	Mean: 2, 4, 7; median: 2, 3, 4; max: 1, 4, 5, 6, 7, 10; min: 1, 2, 3, 6, 7, 8	Mean: 1, 4; std: 1; median: 1; max: 1, 4, 5, 6, 7; min: 3, 8, 9
HFCC	Mean: 1, 2, 3; std: 1; median: 1, 3, 4, 9; max: 2	Mean: 2, 3; median: 3, 9; max: 3; min: 11	Mean: 1; median: 1, 3; max: 1, 2, 3, 4, 5, 6, 7; min: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	Mean: 1, 3, 4, 6, 7, 9, 10; std: 2, 3, 11; median: 1, 2, 3, 4, 6, 8; max: 1, 2, 3, 5, 7; min: 8	Mean: 3; median: 1, 3; max: 1, 2, 3, 4, 10; min: 1, 3, 5, 8, 11	Mean: 1, 2; std: 1; median: 1, 2; max: 1, 2; min: 1, 3
PLP	Mean: 5, 6, 7, 8; median: 6, 10, 11; std: 3, 4, 5, 6, 7, 9; max: 8; min: 8	Mean: 6	Mean: 5, 6, 7, 8, 11; median: 5, 6, 7; max: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	Mean: 4, 5, 9, 13; median: 8, 10, 11, 12, 13; std: 3, 7, 12; max: 6, 12, 13; min: 1, 2, 4, 6, 9, 10, 11, 12	Mean: 1; max: 7	Mean: 5, 13, 14; median: 6, 14; std: 3, 7, 9; max: 7, 8, 9, 10, 13
RPLP	Mean: 1, 2, 3, 7; median: 4, 8; std: 2, 5, 6, 13; max: 2, 4, 9; min: 1	Mean: 1; max: 2, 8, 13; min: 8	Mean: 1, 2, 3, 6, 7; median: 1, 2, 3, 6, 7; std: 1; max: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; min: 2, 3, 6, 7, 8, 9	Mean: 1, 2, 5, 14; median: 1, 2, 3, 5, 6, 7, 9, 14; std: 1, 2, 4, 5, 10, 11, 14; max: 1, 4, 6; min: 1, 2, 3, 11, 13	Mean: 1, 3, 4; median: 2, 5; std: 8; max: 2, 8, 9, 10, 13; min: 1, 2, 3, 6	Mean: 1, 2; median: 2, 3; std: 2; max: 3; min: 6, 8
RASTA PLP	Mean: 1, 2, 4, 6, 9; median: 2, 3, 5, 6; std: 1, 2, 3, 9, 10; max: 1; min: 2, 4, 7	std: 4; max: 1, 10; min: 1, 3, 5, 8	Max: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12; std: 1; min: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	Mean: 4, 9; median: 2, 3, 5, 8; std: 1, 3, 5, 10; max: 6, 9; min: 1, 2, 3, 4, 7	Max: 1, 3, 4, 7; min: 1, 3, 8, 10	std: 1; max: 1, 4, 5, 6; min: 1, 2, 3, 4, 5, 6

data and feature types [26]. In Tables 6 and 7, one can observe that applying SFS had none or very little effect on the quality of classification. From the two selection methods, presented in this paper, the classifier independent FMS proved to be superior in comparison to SFS. The highest classification for *k*-NN were achieved with FMS. Additionally, S-PCA method was used in order to compare results of feature extraction and selection. For SVM, applying S-PCA improved the results far better than the selection methods. However, the best results for SVM were achieved after applying S-PCA on sets of already selected features, as shown in Tables 6 and 7. Using selection first with extraction as a second step helped to reduce the noise from non-discriminant features. Again, FMS gave better results than SFS when combined with S-PCA.

6 Conclusions

For the purpose of the examination, a Polish spontaneous emotion database was created. It consists of over 700 samples divided into seven sets representing primary emotional states. Moreover, for comparative aims, we analyzed emotions performed in Polish, by professional actors.

The main objective of this work is to test the efficiency of perceptual features, used in speech recognition (BFCC, HFCC, RPLP, and RASTA PLP), in emotion recognition. As this research has shown, these features proved to be highly discriminative which justifies their application in emotion recognition.

Acknowledgments
This work has been partially supported by the Estonian Research Grant (PUT638), Estonia-Poland (Est-Pol) Research Collaboration Project (MAJoRA: Multimodal Anger and Joy Recognition by Audiovisual Information), and the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund.

Authors' contributions
DK conceived of the study; participated in the design of the work, data collection, data analysis, interpretation, and coordination; and drafted the manuscript. TS participated in the data collection, data analysis, and interpretation and helped to draft the manuscript. GA participated in the design of the work and critical revision of the article. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Mechatronics and Information Systems, Stefanowskiego 18/22, 90-924 Lodz, Poland. [2]iCV Research Group, Institute of Technology, University of Tartu, Nooruse 1, 50411 Tartu, Estonia. [3]Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep, Turkey.

Received: 19 July 2016 Accepted: 8 February 2017
Published online: 20 February 2017

References

1. R. El Kaliouby, P. Robinson, in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*. Mind reading machines: Automated inference of cognitive mental states from video, vol. 1 (IEEE, 2004), pp. 682–688
2. P. R. De Silva, A. P. Madurapperuma, A. Marasinghe, M. Osano: 2006, in *In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. A multi-agent based interactive system towards child's emotion performances quantified through affective body gestures, vol. 1 (IEEE, pp. 1236-1239
3. D. Kamińska, A. Pelikant, Recognition of human emotion from a speech signal based on Plutchik's model. Int. J. Electron. Telecommun. **58**(2), 165–171 (2012)
4. R. Plutchik, *Emotion: a psychoevolutionary synthesis*. New York Harper and Row, 1980)
5. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, in *Interspeech*. A database of German emotional speech, vol. 5, (2005), pp. 1517–1520
6. D. Kamińska, T. Sapiński, A. Pelikant, in *Signal Processing Symposium (SPS)*. Recognition of emotional states in natural speech, vol. 2013 (IEEE, 2013), pp. 1–4
7. K. Ślot, *Rozpoznawanie Biometryczne Nowe Metody Ilościowej Reprezentacji Obiektów*. (Wydawnictwa Komunikacji i Łączności, 2010)
8. B. W. Schuller, S. Steidl, A. Batliner, in *Interspeech*. The INTERSPEECH 2009 emotion challenge, vol. 2009, (2009), pp. 312–315
9. K. Crammer, Y. Singer, On the algorithmic implementation of multi-class SVMs. JMLR. **2**, 265–292 (2001)
10. D. Ververidis, C. Kotropoulos, in *Proc. Panhellenic Conference on Informatics (PCI)*. A review of emotional speech databases, (Thessaloniki, Greece, 2003), pp. 560–574
11. G. Klasmeyer, Emotions in Speech. Institut fur Kommunikationswissenschaft, Technical University of Berlin (1995)
12. J. Cichosz. Database of polish emotional speech. http//www.eletel.p.lodz.pl/med/eng/. Access Dec 2016
13. S. Mozziconacci, D. Hermes, in *Proceedings of ICPhS99*. Role of intonation patterns in conveying emotion in speech (ICPhS, 1999), pp. 2001–2004
14. C. Busso, S. S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Int. Conf. Acoust. Speech Signal Process. **17**(9), 582–596 (2009)
15. A. Obrębowski, *Narząd Głosu i Jego Znaczenie W Komunikacji Społecznej*. (Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2008)
16. T. Zieliński, *Cyfrowe przetwarzanie sygnałów*. (Wydawnictwa Komunikacji i Łączności, 2013)
17. M. D. Skowronski, J. G. Harris, in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. Increased MFCC filter bandwidth for noise-robust phoneme recognition, vol. 1 (IEEE, 2002), pp. I–801
18. P. Kumar, A. Biswas, A. N. Mishra, M. Chandra, Spoken language identification using hybrid feature extraction methods. J. Telecommun. **1**(2), 11–15 (2010)
19. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J Acoust. Soc. Am. **87**(4), 1738–1752 (1989)
20. H. Hermansky, N. Morgan, RASTA processing of speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1990)
21. Q. Cheng, H. Zhou, J. Cheng, The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **33**(6), 1217–1233 (2011)
22. J. Kittler, Feature set search algorithms. Pattern Recognition Signal Proc. **41**, 60 (1978)
23. H. Z. T. Hastie, R. Tibshirani, Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)
24. A. Abelin, Anger or Fear? — Crosscultural multimodal interpretations of emotional expressions. Emot. Human Voice. **1**, 65–75 (2008)
25. M. Ayadi, M. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**, 572–587 (2011)
26. A. Janecek, W. Gansterer, M. Demel, G. Ecker, On the relationship between feature selection and classification accuracy. JMLR. **4**, 90–105 (2008)