ARCHIVES OF ACOUSTICS
Vol. **46**, No. 2, pp. 259–269 (2021)
DOI: 10.24425/aoa.2021.136580

Research Paper

Acoustic Methods in Identifying Symptoms of Emotional States

Zuzanna PIĄTEK*, Maciej KŁACZYŃSKI

AGH University of Science and Technology
Faculty of Mechanical Engineering and Robotics
Department of Mechanics and Vibroacoustics
Cracow, Poland; e-mail: maciej.klaczynski@agh.edu.pl
*Corresponding Author e-mail: zpiatek@student.agh.edu.pl

(*received April 9, 2020; accepted February 4, 2021*)

The study investigates the use of speech signal to recognise speakers' emotional states. The introduction includes the definition and categorization of emotions, including facial expressions, speech and physiological signals. For the purpose of this work, a proprietary resource of emotionally-marked speech recordings was created. The collected recordings come from the media, including live journalistic broadcasts, which show spontaneous emotional reactions to real-time stimuli. For the purpose of signal speech analysis, a specific script was written in Python. Its algorithm includes the parameterization of speech recordings and determination of features correlated with emotional content in speech. After the parametrization process, data clustering was performed to allows for the grouping of feature vectors for speakers into greater collections which imitate specific emotional states. Using the *t*-Student test for dependent samples, some descriptors were distinguished, which identified significant differences in the values of features between emotional states. Some potential applications for this research were proposed, as well as other development directions for future studies of the topic.

Keywords: emotion recognition; speech signal processing; clustering analysis; Sammon mapping.

1. Introduction

Emotions are an inseparable element of interpersonal interaction. The human voice is one of the natural, spontaneous and authentic means of expressing emotions, which is why the speech signal is used to effectively identify the speaker's emotional states (Igras *et al.*, 2013).

Emotion recognition, dynamic improvements in quality and data processing speed, and the development of the information society, have recently become popular subjects of research, which fit into a wider range of applications, including human-computer communication interfaces (voice assistants), human behaviour monitoring, concentration, and stress level or improving the effectiveness of biometric recognition.

Emotion identification based on voice analysis presents a huge challenge due to the complexity of the speech signal, the main aim of which is content transmission of a given statement. The individual nature of articulation, emotional load, fatigue, and/or medical condition are just a few factors which modify the intricate frequency-time relationships in which this content is contained (Ślot, 2010). Enhancing speech recognition systems (ASR – Automatic Speech Recognition) with the ability to identify speakers' emotions would improve voice interfaces by making interactions with them feel more authentic and adequate.

The aim of this work was to attempt to identify the speaker's emotional state on the basis of a 19-dimensional feature vector obtained as a result of parameterization, to indicate the differences between parameters for two emotional states: calm (neutral) and nervous, and to determine parameters correlated with emotional content in Polish speech. This article discusses the statistical analysis including the Students' *t*-test (selection of descriptors), cluster analysis (grouping of feature vectors by emotion), and Sammon mapping with assessment (Sebestyen's criterion).

2. Analysis of the issue

The main issues in this topic area are the following: choosing a good database of emotional speech recordings, selection of effective features (parameters), and establishing a process for identifying and classifying emotions.

The greatest challenge in recognising emotions is the database of recordings (especially among less popular languages, e.g., Polish) and the parameters of existing and available speech corpora are not always satisfactory. Speech technology systems that include detection of emotions require appropriate research material in the form of a recording database. Having relatively extensive and diverse collection of content and speakers, considering the need also for adequate gender in the emotional speech corpus, is one of the most important and greatest challenges in the analysis of emotional speech.

There are several emotional speech corpora that take into consideration the method of obtaining emotions and the type of labelling and authenticity of emotions. First, *forced emotions* are acquired by recording actors acting out predefined emotions. These are characterised by very good quality recordings and the ability to get a large number of emotions with any content. *Authentic emotions*, where the source of the material are all types of recordings from the media, i.e., live coverage, talk shows, political debates, and recordings from emergency calls. The disadvantages of these sources are usually the poor quality of the recordings, various types of artifacts, noises, or statements of many people at once. On the other hand, these showing authentic emotions recordings present a variety of situations and garner the highest credibility, because the characters' emotions were evoked by real stimuli – sudden situations or life experiences. *Emotions triggered artificially* are acquired using various techniques of inducing emotions, which include emotionally engaging content, like movies, stories, images, or computer games. The drawbacks of this kind of corpus are limitations due to ethical reasons and low intensity or artificiality of experimental situations. Nevertheless, this corpus type offers a good quality of recording and the possibility of properly planning the experiment (Sidorova, 2007; Igras *et al.*, 2013).

Some scientists use ready-made public databases (e.g., the Berlin Database – emotions played), while others decide to prepare corpora suited to the subject of the analysis.

In an article (Ververidis, Kotropoulos, 2003), the authors compiled 32 corpora of emotional speech (11 in English, 8 – German, 3 – Japanese, 2 – Dutch, 2 – Spanish, 1 – Danish, 1 – Hebrew, 1 – Swedish, 1 – Chinese, 1 – Russian, and 1 multilingual), where over 20 of them contained recordings of emotions acted out by actors. Some of the above foreign speech corpora are publicly available, while there are few Polish ones. As a part of the work conducted at the Telecommunications Institute of the Warsaw University of Technology, a database of spontaneous emotions (BES) from Polish Radio and TVP broadcasts was created for the Polish language (Janicki *et al.*, 2008). In turn, Demenko *et al.* (2011) created a database of emergency telephone call recordings, and Cichosz *et al.* (2008) conducted the recording of emotions performed by actors for the purposes of research. In other Polish studies (Kamińska *et al.*, 2012), the research groups used their own recordings, made for the purposes of research, or recordings from the media, collected for analysis. Databases containing recordings exhibiting spontaneous emotions with credible Polish language content are rare. The main problem in creating a large and reliable database is related to the originality of the emotions in the recordings. The expression of emotions by actors and speakers leads to doubts about the validity of the research and its conclusions.

The next stage of automatic recognition is the selection of appropriate features. In general, the set of descriptors commonly used in speech analysis also holds for emotion recognition. Previous studies on the characteristics of the signal correlated with the emotional states of the speaker indicated the following groups of characteristics: the signal energy and parameters describing energy changes (minimum, maximum, range, mean, variance, etc.), the fundamental frequency F_0, waveform parameters of F_0 and their derivative (minimum, maximum, range, mean, variance, etc.), mel-frequency cepstral coefficients (MFCC) coefficients, time-related features (number and length of pauses, speech rate). In recent years, research into the recognition of emotions based on speech has intensified. For example in the article by Davletcharova *et al.* (2015), the authors used a database of recordings containing 4 emotional states acted out in Russian and MFCC parameters were used for the assessment. El Haddad *et al.* (2017) analysed two emotions (smile/laugh and neutral) in two languages (English and French) by first four formants. Zvarevashe and Olugbara (2020) used a combined feature vector (F_0, ZCR, energy, MFCC, LPCC, FFT, Spectral Centroid Moments etc.) to analyse recordings from two databases (RAVDESS – 8 emotions, North American acent and SAVEE – 7 emotions, British English). Other studies worth mentioning are presented in articles (Yeqing *et al.*, 2011; Sun *et al.*, 2015; Razuri *et al.*, 2015; Özseven, 2018; Stolar *et al.*, 2018; Kerkeni *et al.*, 2019; Bhavana *et al.*, 2019; Abdel-Hamid, 2020; Ntalampiras, 2021; Zhang, 2021).

Features related to intonation, stress and period are referred to as prosodic features (Kamińska *et al.*, 2012; Zetterholm, 1998). Based on the collected features, feature vectors are created that are used in the next step – classification. These methods are standard tools but their selection is also an important element.

3. Emotional speech database

Based on the analysis of the aforementioned classes of emotional speech corpora and the availability of databases, a proprietary corpus of recordings from the media was prepared, containing authentic and spontaneous emotions. The first step was to collect speech samples in two emotional states for the same speaker: calm (neutral) and nervous. In live journalistic programs, the feelings or reactions of the participants seem to be spontaneous, provoked by events or discussions, or relating to difficult life situations. Due to the large amount and variety of material on the "publicistykatvp" channel, two programs were selected for presenting content related to engaging political and social issues ("*Tomasz Lis na żywo*"/"*Tomasz Lis Live*" and "*Sprawa dla Reportera*"/"*Reporter's Case*"). All samples were evaluated during material collection and labelled for the above-mentioned states. Then some of the selected recordings were verified by a 3rd-year student of Psychology at the Pedagogical University of Cracow.

The final set of recordings used for further analysis consisted of 48 audio samples from 12 women and 12 men, of different ages, where for each speaker there were 2 samples – in a calm or agitated/anxious state with different content and duration (1–11 seconds).

4. Parameterization

The acoustic signal was modified to form a feature vector, with speech parameters being the basis for the description of emotional states. One of the most important steps is the quantitative description of the subject of research, i.e., the identification of object features that carry information sufficient to effectively identify emotional states.

After collecting the database of emotional speech recordings and subjecting them to a two-stage labelling process, the parameterisation of the samples was carried out. The analysis of speech acoustic signals included the calculation of signal features in the time domain – signal energy, signal power, fundamental frequency F_0 (mean, median, minimum value, maximum value, standard deviation, range), jitter, shimmer – and in the frequency domain – spectral moments (M_0, M_1, M_2), kurtosis, skewness, formants frequencies (F_1, F_2, F_3, F_4).

The purpose of analyzing the obtained results was to show changes in parameter values between the calm and nervous states. Signal parameterisation both in the time domain and in the frequency domain included the calculation of several features for speech signal samples. Among them were the parameters described below.

4.1. Fundamental frequency F_0

The fundamental frequency is an individual feature resulting from the size of the larynx, the tension and size of the vocal cords, depends on gender and age. For example, for men it is in the range of 75–300 Hz and for women in the range of 100–500 Hz (Boersma, Weenink, 2015; 2019). The parameter is indicative of the scale of the voice, and during a conversation the range of changes is associated with intonation, which plays a significant role in the expression of emotions. In this algorithm, the range (considering ranges for both sexes) was set to 75–500 Hz, which means that the method detected the values of the fundamental frequency only in this declared range.

The extraction of the waveforms of the laryngeal tone's fundamental frequency, which is the basic harmonic of the signal reflecting the frequency of the vibrations of the vocal folds, was performed using an algorithm analysing the autocorrelation function. The advantages of the method used were very high resolution and resistance to noise and interference occurring in the signal. From the vector formed of consecutive fundamental frequencies, 6 statistical features were determined such as: mean F_0, median F_0, minimum value F_0, maximum value F_0, standard deviation F_0, range F_0.

4.2. Spectral moments

Before calculating the spectrum, the signal oscillogram was filtered with a high-pass filter – preemphasis. Short-time discrete Fourier transform (STFT) was used due to the speech spectrum changing over time, dividing the signal into 20 ms sections with 10 ms overlaps. Spectral parameters, determined based on signal spectrum estimates, describe the shape and prove to be very useful during analysis. In the described studies, four normalised moments of the m-th order and zero-order moment were determined. The first three spectral moments (zero, first, and second order) were taken into the initial feature vector, while the third and fourth order moments were used only to calculate the remaining kurtosis and skewness parameters. Higher order moments are less useful, even normalised ones, because they are correlated with each other and do not provide a convincing interpretation.

4.3. Formants

Formants are the basic group of parameters used for speech analysis, processing, and recognition by researchers dealing with speech issues.

The vocal tract consists of a series of structures having the ability to vibrate naturally. The larynx tone passing through it is subject to modifications due to natural vibrations of the throat, nasal cavity, or mouth. Thanks to this, certain components of the primary laryngeal tone are strengthened, while others are weakened. Maxima of the amplified frequencies are called formants (Obrębowski, 2008). The values of the formants depend on the individual characteristics (length of the vocal canal) as well as the manner of articulation (degree of rounding of the mouth) (Kamińska *et al.*, 2012).

Usually, vowel formants are set and in the literature can be found clearly declared ranges in which they occur. In the case of continuous speech, subsequent bands with a width of 1000 Hz are analysed in search of the local maximum spectrum and formant frequency for which it occurs. The formant frequency ranges were selected based on the lower and upper frequency values for individual vowels (Kłaczyński, 2007). The first four formants were designated in the described studies.

Based on the speech parameters described above, a 19-dimensional feature vector was created. The above descriptors were calculated for each of the 24 speakers. Table 1 summarises the values of the individual parameters and differences in values between emotions for both male and female speech.

Figures 1 and 2 present comparisons of the oscillogram, spectrogram, and waveform of the fundamental frequency over time in a calm and nervous state.

Table 1. Summary of sample (19-dimensional) feature vectors showing differences in parameter values between the analysed emotional states for female and male speech.

Parameter	Female		Male	
	Calm state	Nervous state	Calm state	Nervous state
F_0 median	248.1	278.1	118.9	271.3
F_0 mean	255.4	300.7	146.6	265.4
F_0 min	81.7	84.3	75.8	132.1
F_0 max	463.2	493.5	499.3	492.7
F_0 std	64.0	105.6	82.7	42.3
F_0 range	381.5	409.2	423.5	360.6
Jitter	5.9	9.3	21.9	6.5
Shimmer	10.8	14.7	15.7	13.9
M_0	0.0046	0.0085	0.0017	0.0100
M_1	5889	6101	5918	6708
M_2	16222423	14900857	18344466	20074062
Kurtosis	2.08	2.35	3.06	3.53
Skewness	$5.85 \cdot 10^{-12}$	$6.43 \cdot 10^{-12}$	$7.48 \cdot 10^{-12}$	$9.35 \cdot 10^{-12}$
F_1	728	731	654	699
F_2	1684	1699	1593	1731
F_3	2844	2904	2941	2981
F_4	3632	3505	3627	3555
Energy	0.95510	1.50348	0.92395	1.87650
Power	0.00108	0.00171	0.00096	0.00196

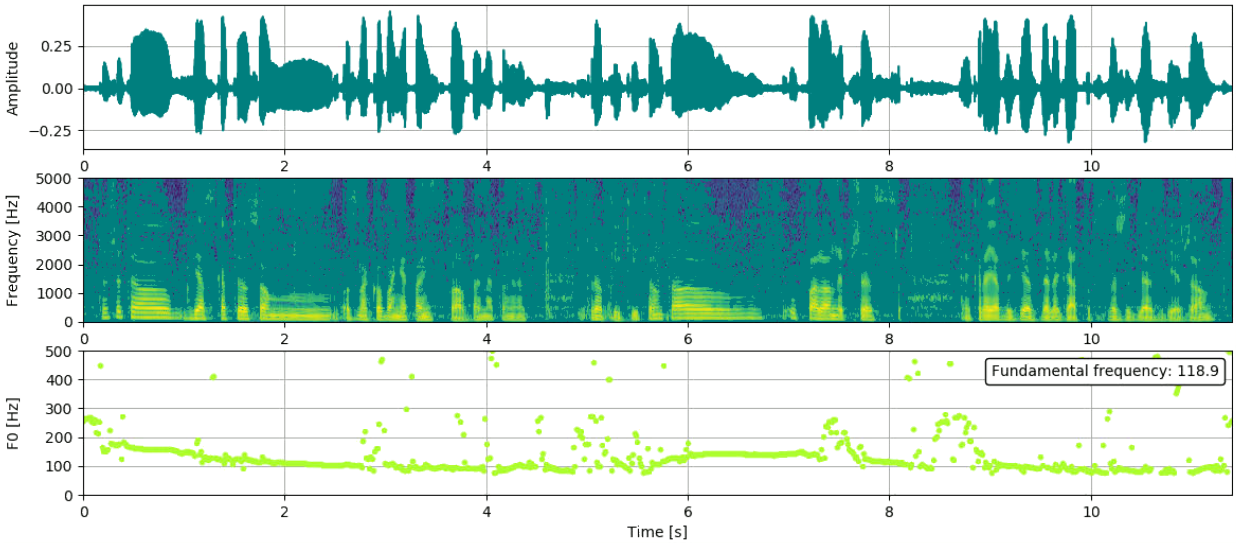

Fig. 1. Time course, spectrogram, and subsequent values of the fundamental frequency for the calm state for male speech.

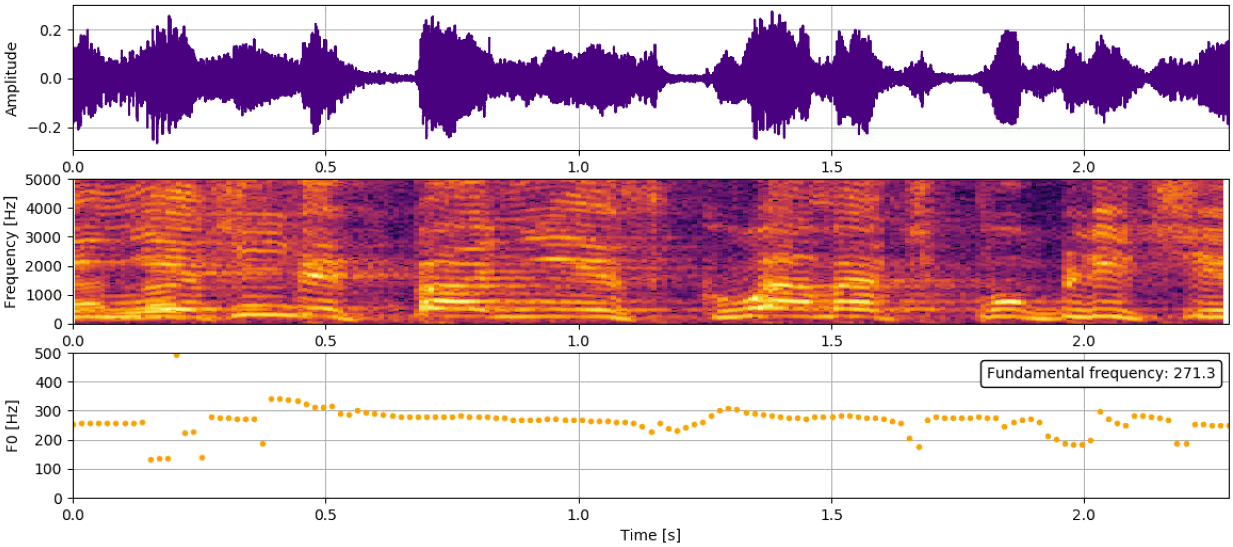

Fig. 2. Time course, spectrogram, and subsequent values of the fundamental frequency for the nervous state for male speech.

Analysis of the deviation waveforms obtained allowed for the observation of intonation variations during the statements for individual emotions and indicated general regularities of parameter changes between the calm and nervous state. For the nervous state, a change in the contour and larger deviations of the fundamental frequency relative to the calm (neutral) state were noted.

5. Feature vectors verification

Research on recognising or identifying emotional states based on a speech signal is based on the assumption of the existence of universal emotional patterns or profiles. In fact, individual differences as well as subjectivity in perception and expression of emotions in speech constitute a great difficulty. An individual way of expressing or perceiving emotions often causes the need to adapt the system to the profile characteristics of a given speaker. Therefore, there is no rule that is always met, which refers to the question whether a given parameter always increases, decreases or does not change significantly in the event of a transition from a calm state to a nervous state for a given speaker.

Verification of the feature vector quality was carried out using three methods: Student's t-test, cluster analysis (Nisbet *et al.*, 2018), and Sammon mapping (Sammon, 1969).

5.1. Student's t-test

The Student's t-test is a frequently used procedure in statistics. For dependent groups, it is applied to compare average values from the same group (for the same parameter) but from different times in this study, a calm versus nervous state of the same speaker and provide information whether the groups differ from themselves and how significant these differences are. In the context of this study, parameter values between states indicate a change in the state of emotions. In order to check how repeatable the results are for a group of speakers, a Student's t-test was used.

Analysing Table 2, the parameters for which there is a statistical difference between the values for emotional states rejecting the null hypothesis consist of: mean F_0, median F_0, standard deviation F_0, M_0, F_1, F_2. An interesting case is the jitter parameter, for which the obtained values between states are different, although the last column suggests that the results may be random – that such a result was obtained for this particular data set. It is not known which values the parameter will have for a different database of recordings and whether jitter will always increase in a nervous state for each test group.

5.2. Cluster analysis

In the classical cluster analysis, a matrix of distances between objects is calculated in order to determine the similarity of these objects. Reversing the assumption, in order to estimate the correctness of feature space choice (the measure of feature dissimilarity), the objects were created from the individual elements of the feature vector (Table 2) taking into account the entire database. In order to calculate the distance between the parameters of the presented feature vector, the Euclidean metric was used and the average method was used for group of objects. The results of this verification are shown in Figs 3 and 4. Before cluster analysis, feature scaling was performed. Feature scaling is a common procedure used in machine learning at the stage of

Table 2. Student t-test results for dependent samples.

Parameters	t-statisics	Critical values	p-value	T_stat *versus* Cv	p-value *versus* alpha
F_0 median	11.319	1.714	0.00000000007	**REJECT – different**	**REJECT – different**
F_0 mean	12.440	1.714	0.00000000001	**REJECT – different**	**REJECT – different**
F_0 min	1.210	1.714	0.23861	ACCEPT – equal	ACCEPT – equal
F_0 max	0.683	1.714	0.50128	ACCEPT – equal	ACCEPT – equal
F_0 std	2.654	1.714	0.01418	**REJECT – different**	**REJECT – different**
F_0 range	0.220	1.714	0.82784	ACCEPT – equal	ACCEPT – equal
Jitter	2.008	1.714	0.05649	**REJECT – different**	ACCEPT – equal
Shimmer	1.182	1.714	0.24919	ACCEPT – equal	ACCEPT – equal
M_0	4.613	1.714	0.00012	**REJECT – different**	**REJECT – different**
M_1	0.690	1.714	0.49724	ACCEPT – equal	ACCEPT – equal
M_2	1.274	1.714	0.21531	ACCEPT – equal	ACCEPT – equal
Kurtosis	0.319	1.714	0.75258	ACCEPT – equal	ACCEPT – equal
Skewness	1.421	1.714	0.16874	ACCEPT – equal	ACCEPT – equal
F_1	2.629	1.714	0.01501	**REJECT – different**	**REJECT – different**
F_2	2.348	1.714	0.02786	**REJECT – different**	**REJECT – different**
F_3	0.571	1.714	0.57379	ACCEPT – equal	ACCEPT – equal
F_4	1.192	1.714	0.24558	ACCEPT – equal	ACCEPT – equal
Energy	0.563	1.714	0.57884	ACCEPT – equal	ACCEPT – equal
Power	0.517	1.714	0.61040	ACCEPT – equal	ACCEPT – equal

appropriate data preparation. Normalisation can be useful and even required when data has input values of different scales. Min-max normalisation was applied to the independent parameters. The results of the parameters' verification are shown in Fig. 3 (dendrogram). It depicts the similarities of the parameters in the feature vector for all speakers in the database for both emotional states. As it can be noticed, the

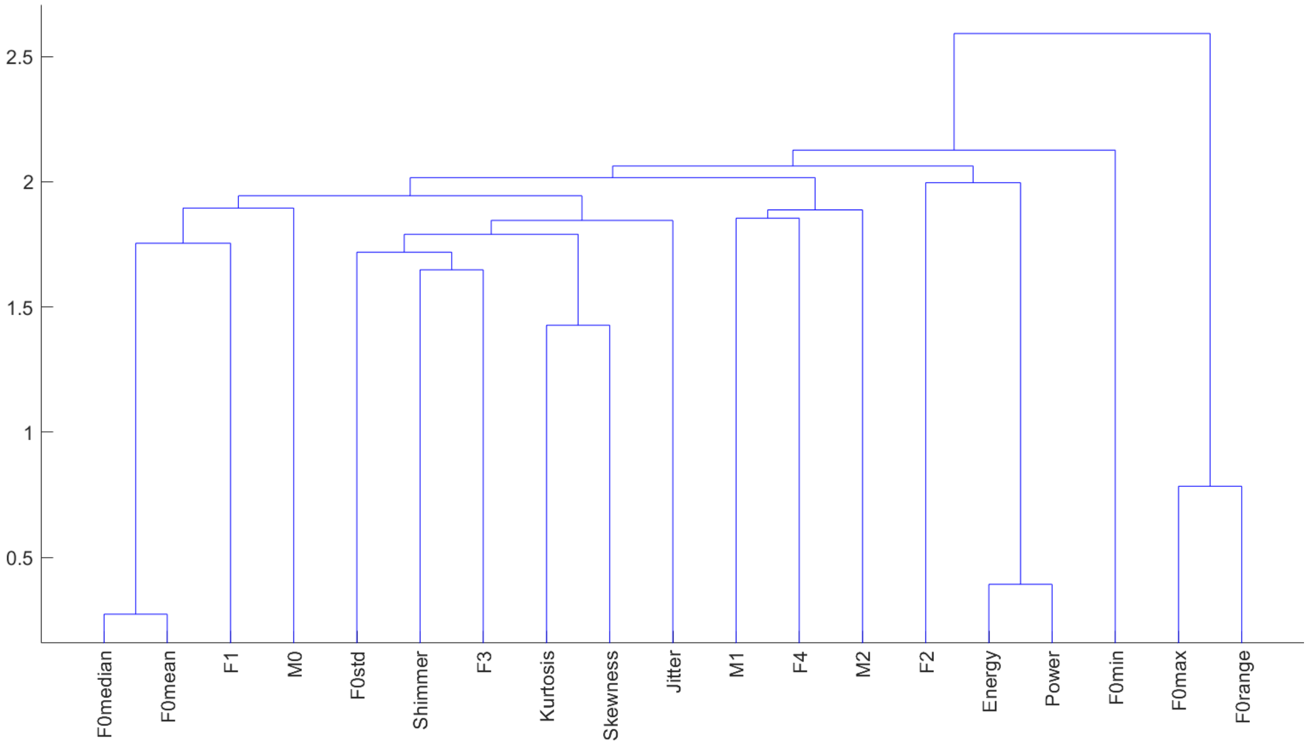

Fig. 3. Results of cluster analysis (dendrogram) showing the similarities between parameters in the feature vector for all speakers in the database for both emotional states.

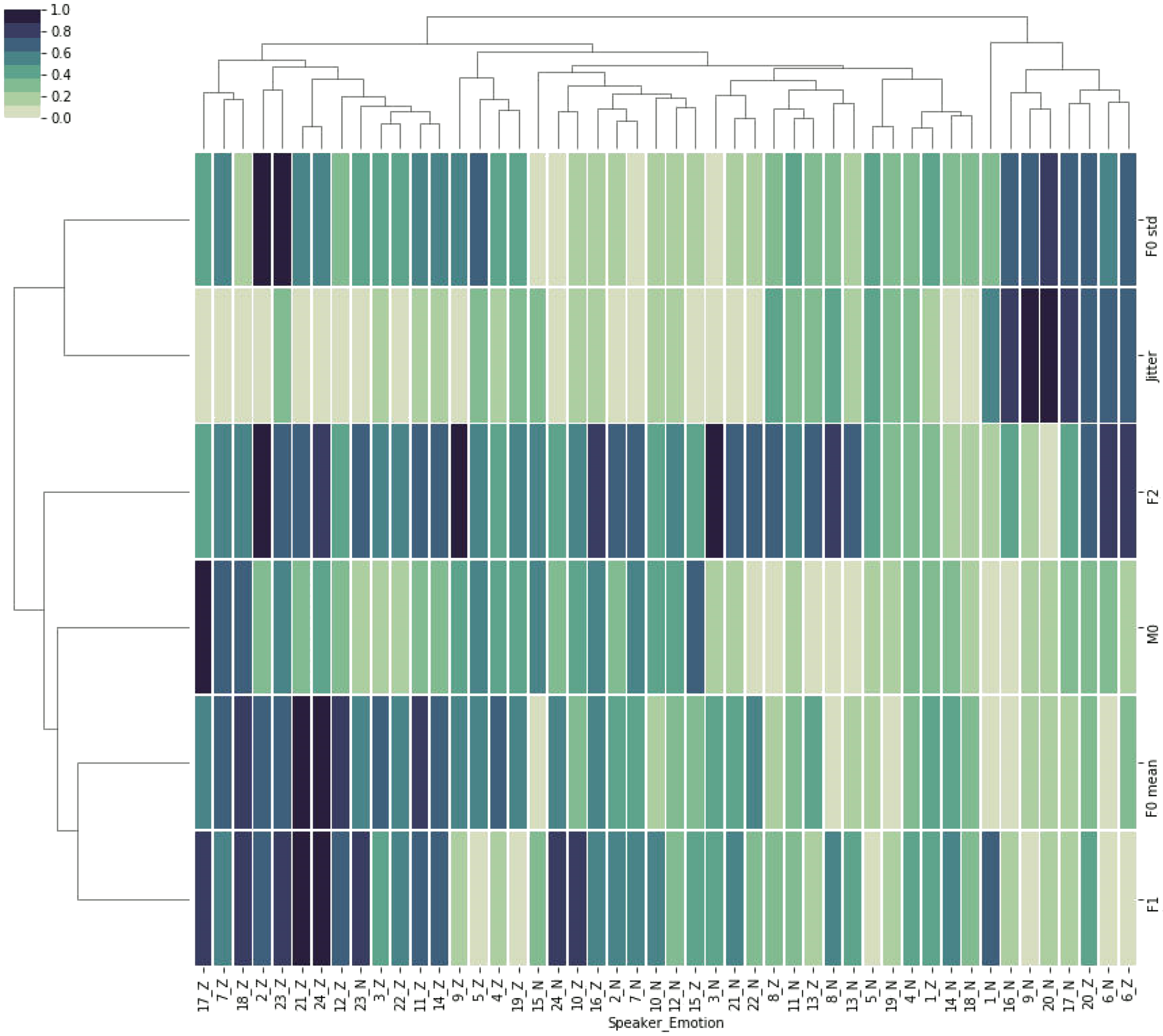

Fig. 4. Cluster analysis, parameter similarities, and feature vectors for a given speaker in terms of emotional state – only for parameters for which a significant difference between groups (emotions) in the t-Student test was indicated.

features which correlated were the following: median F_0 – *mean* F_0; *Energy – Power*; *max* F_0 – *range* F_0. This procedure and the Student's t-test results made it possible to reduce the feature vector to the following parameters: *mean* F_0, *jitter, standard deviation* F_0, M_0, F_1, F_2 for which a double cluster analysis was performed and shown in Fig. 4.

Figure 4 presents a cluster analysis for parameters for which, according to the Student's t-test, there was a significant difference in the value of parameters between emotional states, as well as the second cluster analysis for checking the clusters of the speakers' statements in two emotional states. Limiting the feature vector only to those accurately describing the object of analysis clusters has enabled the recordings to be combined into larger classes, bringing together a greater number of samples of a given emotional state, e.g. from the left, [15_N – 11_N] (6 samples), [17_Z – 23_Z] (13 samples), and based on the value vector, they are classified into a particular emotion. Thus, the performed reduction of the feature space, based on the result of cluster analysis, shows a good prognosis for the possibility of distinguishing the emotional state on the basis of Polish speech signals. However, full validation of these conclusions was made by performing Sammon mapping.

5.3. Sammon mapping

Sammon mapping (Sammon, 1969) allows for the presentation of multivariate data (N-dimensional) on a plane (space R^2). Due to the fact that multidimensional structures are not subject to human interpretation, human ability to imagine the location of these data points in N-dimensional space is futile. There is, therefore, a problem to create an understandable and accessible graphical representation of this data type (Kłaczyński, 2007).

Sammon mapping is a non-linear mapping of points from the R^N space onto their respective "projections", i.e., lying points in R^2 space. For the purposes of this study, 19-dimensional vectors (parameterised signals of calm and nervous state speech) were mapped onto a plane, acting as an implementation of a self-organising learning process. The Sebestyen criterion was used to assess the mapping of the parameterised 19-dimensional feature vectors on the plane. This criterion is based on the measure of intra-class dispersion and global interclass dispersion. To assess the quality of the entire set of N features, the criterion being the logarithm of the dispersion ratio of interclass and intra-class scattering (Kłaczyński, 2007) can be used. Sammon mapping was done for three cases of the feature vector:

1) 19-dimensional feature vector,
2) 7-dimensional feature vector (*mean* F_0, *median* F_0, *standard deviation* F_0, *jitter*, M_0, F_1, F_2),
3) 6-dimensional feature vector (without the jitter parameter).

As shown by the presented results of measures in Tables 3–5, mapping of multidimensional space to a plane is not a real reflection of the original data set (which was anticipated). However, it retains such features as inter-class and intra-class recognition in fairly high compliance with the original vectors. By visually analysing the mapping results for 19 speech signal parameters (Fig. 5) it should be stated that it is not possible to recognise the speaker's emotional state. The points on the plane representing the two states (neutral and nervous) are mixed together. However, when using a vector with 7 parameters, there is a clear improvement in grouping emotional states (Fig. 6). Interestingly, withdrawing the jitter parameter (Fig. 7) did not significantly help to clarify the dividing line between states.

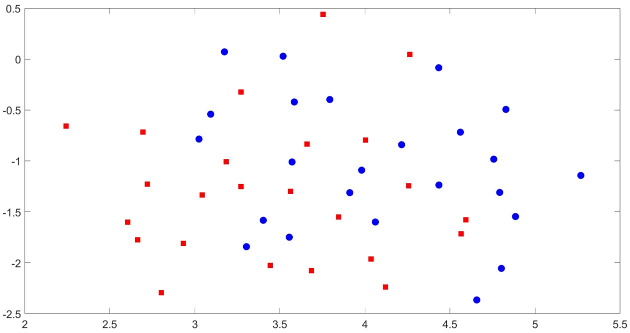

Fig. 5. Sammon mapping visualization (circle – neutral state, square – nervous state) – for the 19 dimensional feature vector.

Table 3. Summary of Sebestyen measures for pre and post Sammon mapping (Fig. 5).

Data	M1 – measure of inter-class dispersion	M2 – measure of intra-class dispersion	Quantitative assessment of class separability $\log_2(\text{M1}/\text{M2})$
Original	2.20	1.84	0.26
After Sammon mapping	2.15	1.73	0.31

Table 4. Summary of Sebestyen measures pre and post Sammon mapping (Fig. 6).

Data	M1 – measure of inter-class dispersion	M2 – measure of intra-class dispersion	Quantitative assessment of class separability $\log_2(\text{M1}/\text{M2})$
Original	0.98	0.64	0.61
After Sammon mapping	1.06	0.57	0.89

Table 5. Summary of Sebestyen measures for pre and post Sammon mapping (Fig. 7).

Data	M1 – measure of inter-class dispersion	M2 – measure of intra-class dispersion	Quantitative assessment of class separability $\log_2(\text{M1}/\text{M2})$
Original	0.84	0.51	0.71
After Sammon mapping	0.90	0.46	0.97

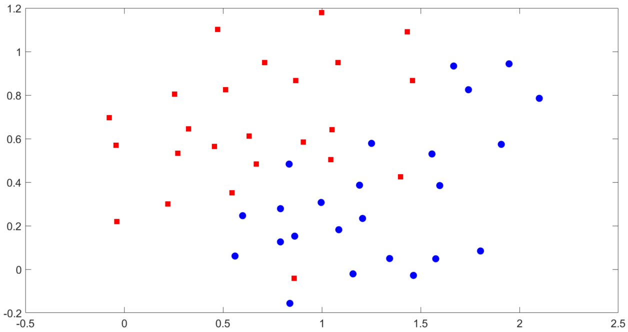

Fig. 6. Sammon mapping visualization (circle – neutral state, square – nervous state) – for a 7 dimensional feature vector.

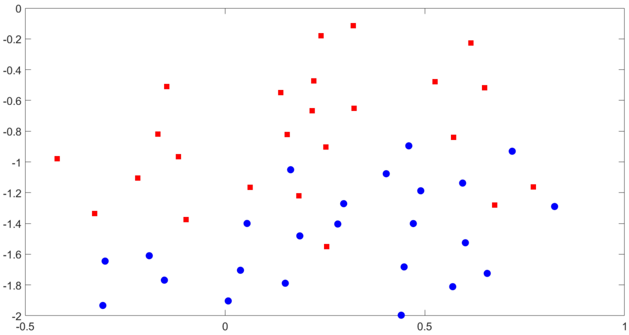

Fig. 7. Sammon mapping visualization (circle – neutral state, square – nervous state) – for a 6 dimensional feature vector.

6. Summary

The main aims of the presented research were to identify differences in parameter values between the calm (neutral) state and the anxious state and to identify these emotional states on the basis of feature vectors for a given speaker. For this purpose, all the necessary functions were implemented in Python, parameterisation of recordings was performed, presenting the tables with the aggregated results above. A Student's t-test for dependent samples was carried out to indicate a statistical difference between two groups (values of a given parameter for speakers between a calm and nervous state). The t-Student test allowed for the extraction of those features that showed significant changes in parameters between emotions, and consequently reduced the feature vector, discarding those that did not show significant changes. Cluster analysis was also performed on the set, showing similarities between parameters, as well as entire feature vectors for a given speaker. The inference was supported by Sammon mapping which confirmed the legitimacy of reducing attributes from the adopted initial vector from 19 to 7 features.

A literature review showed that many studies have managed to achieve satisfactory accuracy and efficiency in emotion recognition using other databases, parameters, or classification methods than those indicated in the publication.

For example, in (Razuri *et al.*, 2015) regarding an English database, to reduce the size of the speech feature vector and improve the results obtained by the classifiers, the output data from a decision tree classifier like feature selection method was used. The decision tree has obtained the most accurate result and was composed of six nodes (six features of the dataset). This meant that the tree only needed these six features to predict the emotions. These results showed that the feature vector was reduced not only to extract those features that show significant changes but also to obtain better efficiency and accuracy.

In (Kerkeni *et al.*, 2019), the authors presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions based on two different acted databases (Spanish and Berlin). The authors studied how classifiers and features (MFCC and MS), as well as their selection, impact the recognition accuracy of emotions in speech. Feature selection techniques showed that more information was not always good in machine learning applications. The results showed that the SER achieved the highest recognition rate of 94% on the Spanish database using the RNN classifier, and for the Berlin database, all of the classifiers achieve an accuracy of 83%. RNN often performs better with more data, however, is limited by very long training times. Therefore, the authors concluded that the SVM and MLR models have good potential for practical usage on limited data in comparison with RNN.

Zvarevashe and Olugbara (2020) proposed the combination of prosodic and spectral features from a group of selected features to realize hybrid acoustic features for improving the efficiency of emotion recognition. The proposed set of acoustic features proved highly accurate in recognising all the eight emotions investigated in the study. The researchers used two public acted speech databases to train five popular ensemble learning algorithms. Results showed that random decision forest ensemble learning of the proposed hybrid acoustic features was highly effective for speech emotion recognition. The achievement high precision when recognising the neutral emotion by using either pure MFCC features or a combination of MFCC, ZCR, energy, and fundamental frequency features (that were seen to be effective in recognising the surprise emotion) proved to be difficult.

Studies on speech emotion recognition have both academic and practical significance. In addition to numerous prospective technological applications (including as a module of automatic speech and speaker recognition systems), applications built on the basis of algorithms that identify the patient's emotional state based on the acoustic parameters of his or her speech have great potential in the field of diagnostics and medical therapy (Igras *et al.*, 2013). In medicine, the applications include the diagnosis of psychological and neurological disorders, the symptoms of which are both abnormal perception and expression of emotions, which include stress (contributing to many diseases of civilization), depression, schizophrenia, and autism (Obrębowski, 2008).

Specific groups of people for whom identifying extreme emotional states may be particularly important can be specified, e.g., pilots or players. In the case of pilots, developing a system that registers the pilot's statements in real time could be useful, so that after performing parameterisation and classification, it will return his emotional state. Moreover, it would have the capacity to provide evaluation, e.g., send information to the military base when the pilot is in a nervous or angered state that could endanger his or her life as well as the lives of others.

An important problem is also the amount of violence and profanity, especially verbal, in the case of streaming games. Children are the most vulnerable, therefore, they may face numerous psychological problems. The solution in this case would be a system that monitors players' reactions and their communication during gaming. If an agitated state is detected leading to misbehaviour, for example, when a player offends or intimidates others, such a system could automatically block players from the games' current or next round. The proposed solution could raise players' awareness of acceptable inter-personal behaviour on the Internet and perhaps protect the youngest players.

References

1. Abdel-Hamid L. (2020), Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features, *Speech Communication*, **122**: 19–30, doi: 10.1016/j.specom.2020.04.005.
2. Bhavana A., Chauhanb P., Rajiv H., Shahc R. (2019), Bagged support vector machines for emotion recognition from speech, *Knowledge-Based Systems*, **184**: 104886, 1–7, doi: 10.1016/j.knosys.2019.104886.
3. Boersma P., Weenink D. (2015–2019), *Praat documentation – Manual*, from http://www.praat.org/.
4. Cichosz J. (2008), *The use of selected speech signal features to recognize and model emotions for the Polish language* [in Polish: *Wykorzystanie wybranych cech sygnału mowy do rozpoznawania i modelowania emocji dla języka polskiego*], Ph.D. Thesis, Lodz University of Technology, Łódź.
5. Davletcharova A., Sugathan S., Abraham B., James A.P. (2015), Detection and analysis of emotion from speech signals, *Procedia Computer Science*, **58**: 91–96, doi: 10.1016/j.procs.2015.08.032.
6. Demenko G., Jastrzębska M. (2011), Analysis of voice stress in emergency calls, [in Polish: Analiza stresu głosowego w rozmowach z telefonu alarmowego], *XVIII Conference on Acoustic and Biomedical Engineering 2011*, Zakopane.
7. El Haddad K. *et al.* (2017), Introducing AmuS: The Amused Speech Database, *Proceedings of 5th International Conference on Statistical Language and Speech Processing SLSP 2017At: Le Mans*, France, pp. 229–240, doi: 10.1007/978-3-319-68456-7_19.
8. Igras M., Ziółko B. (2013), Database of emotional speech recording, *Studia Informatica*, **34**(2B): 67–77.
9. Janicki A., Turkot M. (2008), Recognition of the speaker's emotional state using the support vector machine (SVM) [in Polish: Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających (SVM)], *Przegląd Telekomunikacyjny – wiadomości telekomunikacyjne*, **2008**(8–9): 994–1005.
10. Kamińska D., Pelikant A. (2012), Spontaneus emotion redognition from speech signal using multimodal classification [in Polish: Zastosowanie multimodalnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej], *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, **3**: 36–39.
11. Kerkeni L. *et al.* (2019), Automatic speech emotion recognition using machine learning, [in:] *Social Media and Machine Learning*, doi: 10.5772/intechopen.84856.
12. Kłaczyński M. (2007), *Vibroacoustic phenomena in the human voice channel* [in Polish: *Zjawiska wibroakustyczne w kanale głosowym człowieka*], Ph.D. Thesis, AGH University of Science and Technology, Kraków.
13. Nisbet R., Miner G., Yale K. (2018), *Handbook of Statistical Analysis and Data Mining Applications*, 2nd ed., Elsevier, doi: 10.1016/C2012-0-06451-4.
14. Ntalampiras S. (2021), Speech emotion recognition via learning analogies, *Pattern Recognition Letters*, **144**: 21–26, doi: 10.1016/j.patrec.2021.01.018.
15. Obrębowski A. (2008), *Voice organ and its importance in social communication*, [in Polish: *Narząd głosu i jego znaczenie w komunikacji społecznej*], Publisher University of Medical Sciences, Poznan.
16. Özseven T. (2018), Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition, *Applied Acoustics*, **142**: 70–77, doi: 10.1016/j.apacoust.2018.08.003.
17. Razuri J.G. *et al.* (2015), Speech emotion recognition in emotional feedback for Human-Robot Interaction, *International Journal of Advanced Research in Artificial Intelligence*, **4**(2): 20–27, doi: 10.14569/IJARAI.2015.040204.
18. Sammon J. (1969), A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, **C-18**(5): 401–409, doi: 10.1109/T-C.1969.222678.
19. Sidorova J. (2007), *Speech Emotion Recognition*, Master Thesis, Universitat Pompeu Fabra, Barcelona, doi: 10.13140/RG.2.1.3498.0724.

20. Ślot K. (2010), *Biometric recognition. New methods for the quantitative representation of objects* [in Polish: *Rozpoznawanie biometryczne. Nowe metody ilościowej reprezentacji obiektów*], WKŁ, Warszawa.

21. Stolar M. *et al.* (2018), Acoustic characteristics of emotional speech using spectrogram image classification, [in:] *12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–5, doi: 10.1109/ICSPCS.2018.8631752.

22. Sun Y., Wen G., Wang J. (2015), Weighted spectral features based on local Hu moments for speech emotion recognition, *Biomedical Signal Processing and Control*, **18**: 80–90, doi: 10.1016/j.bspc.2014.10.008/.

23. Ververidis D., Kotropoulos C. (2003), A review of emotional speech databases, *9th Panhellenic Conference on Informatics (PCI)*, Thessaloniki, Greece, http://delab.csd.auth.gr/bci1/Panhellenic/560ververidis.pdf.

24. Yeqing Y., Tao T. (2011), An new speech recognition method based on prosodic analysis and SVM in Zhuang language, [in:] *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, pp. 1209–1212, doi: 10.1109/MEC.2011.6025684.

25. Zetterholm E. (1998), Prosody and voice quality in the expression of emotions, [in:] *Proceedings of 7th Australian International Conference on Speech Science and Technology*, pp. 109–113, Australian Speech Science and Technology Association, Sydney.

26. Zhang Z. (2021), Speech feature selection and emotion recognition based on weighted binary cuckoo search, *Alexandria Engineering Journal*, **60**(1): 1499–1507, doi: 10.1016/j.aej.2020.11.004.

27. Zvarevashe K., Olugbara O. (2020), Ensemble learning of hybrid acoustic features for speech emotion recognition, *Algorithms*, **3**(3), 70, doi: 10.3390/a13030070.