ARCHIVES OF ACOUSTICS
Vol. **46**, No. 2, pp. 271–277 (2021)
DOI: 10.24425/aoa.2021.136581

Research Paper

Audio Feature Space Analysis for Emotion Recognition from Spoken Sentences

Lukasz SMIETANKA*, Tomasz MAKA

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology
Szczecin, Poland; e-mail: tmaka@wi.zut.edu.pl
*Corresponding Author e-mail: lsmietanka1892@gmail.com

(*received May 25, 2020; accepted November 25, 2020*)

An analysis of low-level feature space for emotion recognition from the speech is presented. The main goal was to determine how the statistical properties computed from contours of low-level features influence the emotion recognition from speech signals. We have conducted several experiments to reduce and tune our initial feature set and to configure the classification stage. In the process of analysis of the audio feature space, we have employed the univariate feature selection using the chi-squared test. Then, in the first stage of classification, a default set of parameters was selected for every classifier. For the classifier that obtained the best results with the default settings, the hyperparameter tuning using cross-validation was exploited. In the result, we compared the classification results for two different languages to find out the difference between emotional states expressed in spoken sentences. The results show that from an initial feature set containing 3198 attributes we have obtained the dimensionality reduction about 80% using feature selection algorithm. The most dominant attributes selected at this stage based on the mel and bark frequency scales filterbanks with its variability described mainly by variance, median absolute deviation and standard and average deviations. Finally, the classification accuracy using tuned SVM classifier was equal to 72.5% and 88.27% for emotional spoken sentences in Polish and German languages, respectively.

Keywords: speech analysis; classification; emotional speech.

1. Introduction

Humans show various emotions in their day to day life in different situations. The ability to evaluate an emotional state of the speaker could be the crucial element in interpersonal communication. Such skill allows getting information about the feelings of the interlocutor. This information can be obtained from facial expressions, speech or physiological signals. The use of speech signal to detect the emotional state seems to be the easiest way due to simple acquisition conditions and the fact the speech is a fundamental form of human communication.

The speaker's emotional state can improve the process of interaction in dialogue systems and has a broad range of applications in voice-based human-computer interaction systems. Since such a state has a close connection with the personality of the speaker, it can be used as a part of multimodal biometric systems. The widespread application of the speech-based emotional analysis may be the detection of intoxication of sleepless states of the speaker.

The problem of emotion recognition from speech signal can be viewed as a typical problem of data classification where the most important role plays the representation of speech in the feature space. Therefore, despite many existing systems for this task, the selection of the acoustic features is not apparent. It highly depends on the speaker characteristics and acquisition conditions. However, as many studies show, the linguistic information, along with acoustic information, can improve the overall accuracy of such systems. The emotional properties of voice are derived from attributes such as pitch, loudness, length and vowel quality – stressed parts of spoken sentence influences on the temporal characteristics. The period and the loudness of the sentence regions depend on the stressed and unstressed patterns. Therefore, the variability of the audio features in the time should be included in the analysis stage.

The way how human express their emotions depends mainly on its anatomical properties, age, gender, environment and situational context. Therefore, designing a robust system for emotion recognition using speech signals is a demanding task. In the last decade, many solutions for emotional state recognition from acoustics signals have been proposed and can be found in the literature.

Recently, an approach using gammatone frequency cepstral coefficients was presented. In the experimental stage, two emotional databases were used, German and Chinese. The study conducted using two classifiers, K-nearest neighbours and long short term memory network. In both cases, the obtained best results exceeded 90% (Zhu, Ahmad, 2019). Another study (Feraru, Zbancioc, 2013) focused on the emotion recognition used LPC features. The proposed system uses W-KNN classifier and various combinations of speech signal features, among others such as LPCC, MFCC, LAR, PARCOR, fundamental frequency and formants. In the result, it was shown that using an extended set of the LPC was not introduced more information to improve the performance in emotion the recognition which remains around 90%. Kathiresan and Dellwo (2019) proposed cepstral derivatives in MFCCs for emotion recognition. In the experimental stage, two emotional databases were used with GMM classifier. The results show that the MFCCs used with feature cepstral delta and delta-delta can improve the performance of specific emotions such as *boredom*. However, the overall classification accuracy for both used datasets was respectively 67.5% and 60.8%. Another work (Hao *et al.*, 2019) shows the use of the SVM classifier based on sequential minimal optimization (SMO). The research used speech signal features, among others such as energy, pitch and MFCC. These features were determined for two databases, German and Chinese. In both cases, the results obtained were about 80%. Moreover, with the rapid development of deep learning, many approaches use such technique in emotion recognition. For example, in (Rajak, Mall, 2019), the authors used a convolutional neural network. The research was carried out using one dataset with 1440 samples and MFCC coefficients. The results obtained were around 50–55%. Another study using deep neural network is (Lee *et al.*, 2019) in which one base of emotional speech was used. The used features included MFCC with derivatives and fundamental frequency. Using deep neural network, the highest score (69.4%) was obtained for MFCC without derivatives. An example of using deep learning network is presented in (Meng *et al.*, 2019) where two emotional speech datasets were used. The classification accuracy obtained for the log-Mel spectra of speech as feature space was around 90%.

Since the acoustic feature space plays a crucial role in the classification process, there were many various features has been proposed for emotional state recognition from a speech signal. The extensive analysis of feature sets used for this task can be found in (Ververidis, Kotropoulos, 2006; Anagnostopoulos *et al.*, 2015; Swain *et al.*, 2018). In this work, we have analysed a set of popular low-level audio features which map time-frequency properties differently along with their temporal variability for speech utterances.

Our primary motivation behind the study was to determine the most influential low-level audio attributes extracted from speech waveform on the emotional speech utterances classification. The physiological properties of the speech signal are highly dependent on many factors, including the speaker's gender, age, language, voice pathologies, etc. It was interesting to find out how popular feature sets may have an impact of distinguishing between the types of emotional states in the spoken sentences. The main contribution of this work is an extensive analysis of large feature space of acoustical features and its comparison for two different databases with emotionally tinged utterances in Polish and German languages.

The rest of this paper is organized as follows: Sec. 2 presents audio datasets used in the experimentation phase. An analysis of discriminative properties of audio feature space is described in Sec. 3. The Sec. 4 contains a description of obtained results and its influence on the classification accuracy of emotional states. The last section concludes the whole study.

2. Speech data

In our study, we have used two databases of emotional speech. The first data set (dbDE) is called Emo-DB (Burkhardt *et al.*, 2005) and it contains 535 sentences in the German language recorded as monophonic with sample rate equal to 16 kHz. The utterances are spoken by ten professional actors including five men at the age of 25, 26, 30, 31, and 32 years old as well as five women at the age of 21, 31, 32, 34, and 35 years old. These actors speak ten different sentences lasting from 1 to 5 seconds in the following emotional states: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *boredom*, and *neutral*. The second dataset (dbPL) used in experiments is called Database of Polish Emotional Speech (Slot *et al.*, 2009). It contains 240 examples of emotional speech in the Polish language recorded as monophonic with 44.1 kHz sampling rate. The utterances by four men and women actors were recorded, where every actor spoke five sentences in the emotional states as *anger*, *boredom*, *fear*, *happiness*, *sadness* and *neutral*. Recordings representing *disgust* emotional state in Emo-DB were omitted to obtain the same set of emotions in both databases due to perform consistent comparisons.

3. Feature space analysis

Since the audio features are the crucial part of any audio recognition system, in this study, we have analyzed many popular features and its statistical properties for emotional speech (Mitrovic *et al.*, 2010; Eyben, 2016). The parametrization stage was performed using 25 ms frames of speech with overlapping equal to 10 ms applying for each frame the Hamming window and pre-emphasis filtering. The proposed feature set contains many attributes computed in time and frequency domain. The feature space includes the energy of the signal, fundamental frequency (F0) (Boersma, Paul, 1993; Boersma, Weenink, 2001), linear prediction coefficients (LPC) (Markel, Gray, 1976), linear predictive cepstral coefficients (LPCC) (Rao *et al.*, 2015), Mel frequency cepstral coefficients (MFCC) (Davis, Mermelstein, 1980), and bark frequency cepstral coefficients (BFCC) (Kuan *et al.*, 2016). The selection of fundamental frequency for the whole spoken sentence seems to be the most promising part of the feature space. It is because F0 trajectory represents the properties of the vocal tract; it carries information related to the speaker and the prosody attributes such as intonation and rhythm. For this reason, those properties should provide much information about the characteristics of emotional expression of voice. To include dynamic information of speech in our feature set, we decided to use the velocity and acceleration of the specific attribute changes over time. For this purpose, we computed the delta (Δ) and double delta (ΔΔ) trajectories and included them to the final set of features.

For every feature from the set, a contour was calculated as an attribute value computed for consecutive frames of the analysed signal. Then, for each contour, a 13th-dimensional vector was computer containing 13 statistical properties such as lowest (MIN) and highest (MAX) value, range (RNG), mean (MEAN), standard deviation (STD), first quartile (Q1), median (ME), third quartile (Q3), interquartile range (IQR), quartile deviation (QD), average deviation (AD), median absolute deviation (MAD) and variance (VAR). Table 1 shows the complete list of the proposed features used in the experiments with the dimensionality. The total number of features in the initial set is equal to 3198. In the next step, the features selection procedure was conducted. We have used the SelectKBest algorithm (Pedregosa *et al.*, 2011) with χ^2 test as a scoring function. The algorithm selects the best features based on univariate statistical tests. The Tables 2 and 3 show features sets achieved in the result of the selection process for both datasets. In the first case, for the dbPL, the dimensionality was reduced by 80% and by 79% for the dbDE dataset. According to Table 2, none of properties of double delta trajectories have been included in the feature set after selection process. It can be probably connected with the prosodic attributes of language like melody or rhythm. In order to determine the quality of obtained feature sets, we have employed three classifiers representing different types of classifications. The first of these is Support Vector Machine (SVM), whose classification process is an attempt to separate individual classes using a hyperplane (Chang, Lin, 2011). The next is Random Forest (RF) based on a large number of individual decision trees that operate as an ensemble and each tree in a random forest throws out the class forecast, and the class with the most votes becomes model's forecast (Breiman, 2001). The last one is Naive Bayes (NB) which is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features (Zhang, 2004).

The classification results for the feature sets and selected classifiers are shown in Table 4 and Fig. 1 presents the participation of statistics used for the final feature space. In addition, Fig. 2 shows the contribution of the derivatives (Δ, ΔΔ) in the final feature

Table 1. Initial set of audio features used in the experiments.

Features	Label	Dimensionality
EN + Δ + ΔΔ	EN	39
F0 + Δ + ΔΔ	F0	39
LPC20 + Δ + ΔΔ	LPC	780
LPCC20 + Δ + ΔΔ	LPCC	780
MFCC20 + Δ + ΔΔ	MFCC	780
BFCC20 + Δ + ΔΔ	BFCC	780
Total		3198

Table 2. List of features obtained in the feature selection process for dbPL dataset.

Features	Dimensionality
EN + Δ	25
F0 + Δ	10
LPC13 + Δ	314
LPCC4 + Δ	101
MFCC5 + Δ	99
BFCC11 + Δ	105
Total	654

Table 3. List of features obtained in the feature selection process for dbDE dataset.

Features	Dimensionality
EN + Δ + ΔΔ	20
F0 + Δ + ΔΔ	32
LPC15	189
LPCC6	69
MFCC15 + ΔΔ	158
BFCC20 + ΔΔ	204
Total	672

Table 4. Classification accuracy results using reduced feature sets for selected classifiers and both datasets.

Classifier	Energy		F0		LPC		LPCC		MFCC		BFCC	
	dbPL	dbDE	dbPL	dbDE	dbPL	dbDE	dbPL	dbDE	dbPL	dbDE	dbPL	dbDE
SVM	52.50	58.02	46.25	59.25	46.25	67.90	48.75	62.34	72.50	73.45	58.75	83.33
NB	40.00	44.44	27.50	46.76	31.25	40.12	42.50	28.39	58.75	48.76	47.50	67.28
RF	43.75	50.61	43.75	50.61	37.50	47.53	43.75	50.00	43.75	54.32	42.50	69.13

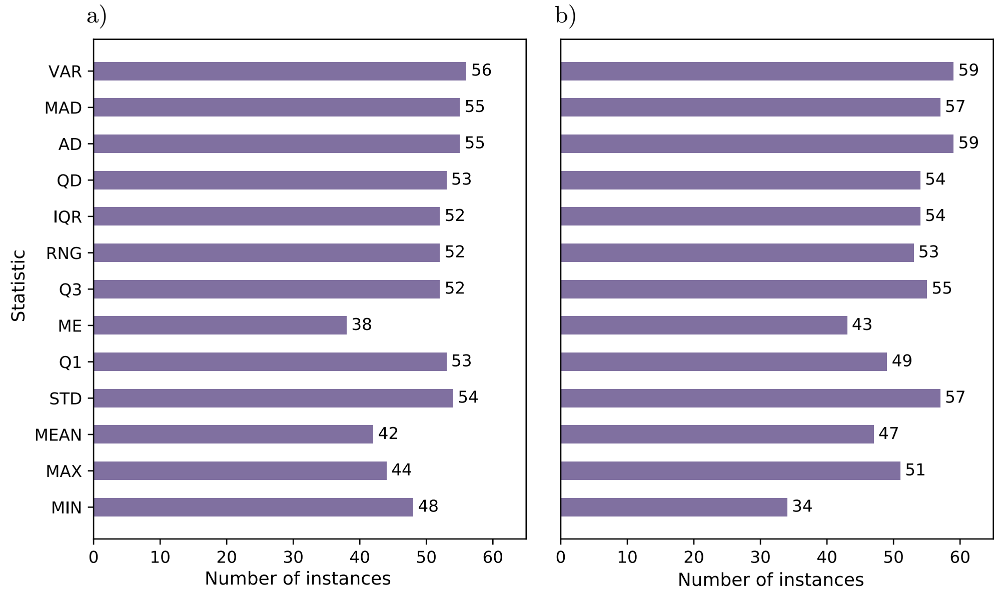

Fig. 1. The number of statistics instances used to compute the reduced feature set from feature contours for dbPL (a) and dbDE (b) datasets.

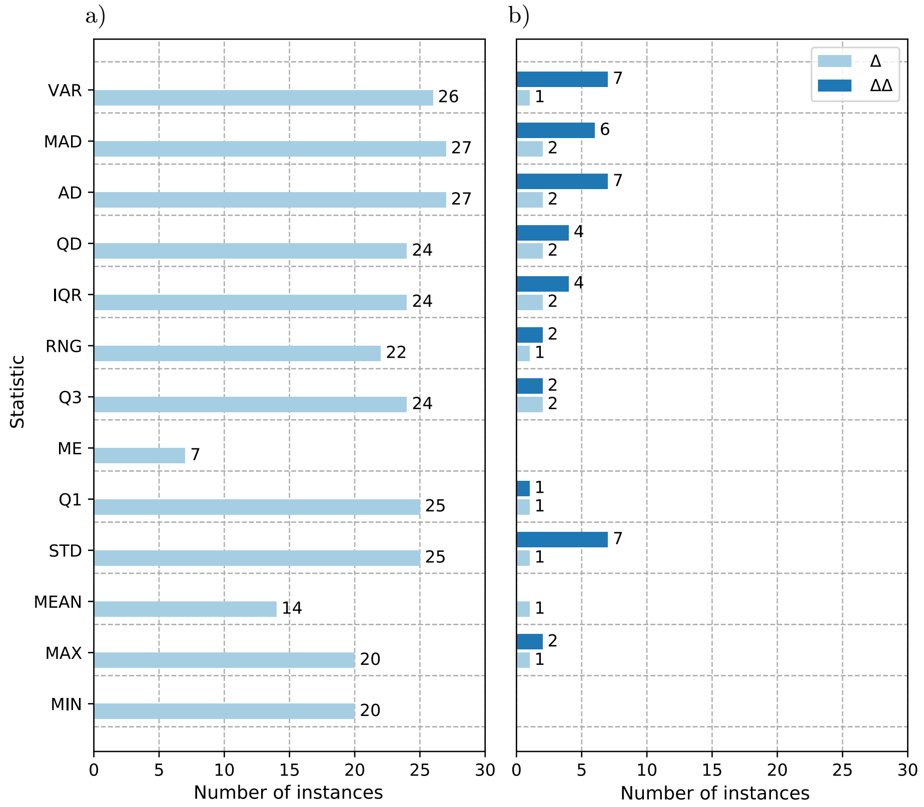

Fig. 2. The number of statistics instances based on derivatives (Δ, ΔΔ) used to compute the reduced feature set from feature contours for dbPL (a) and dbDE (b) datasets.

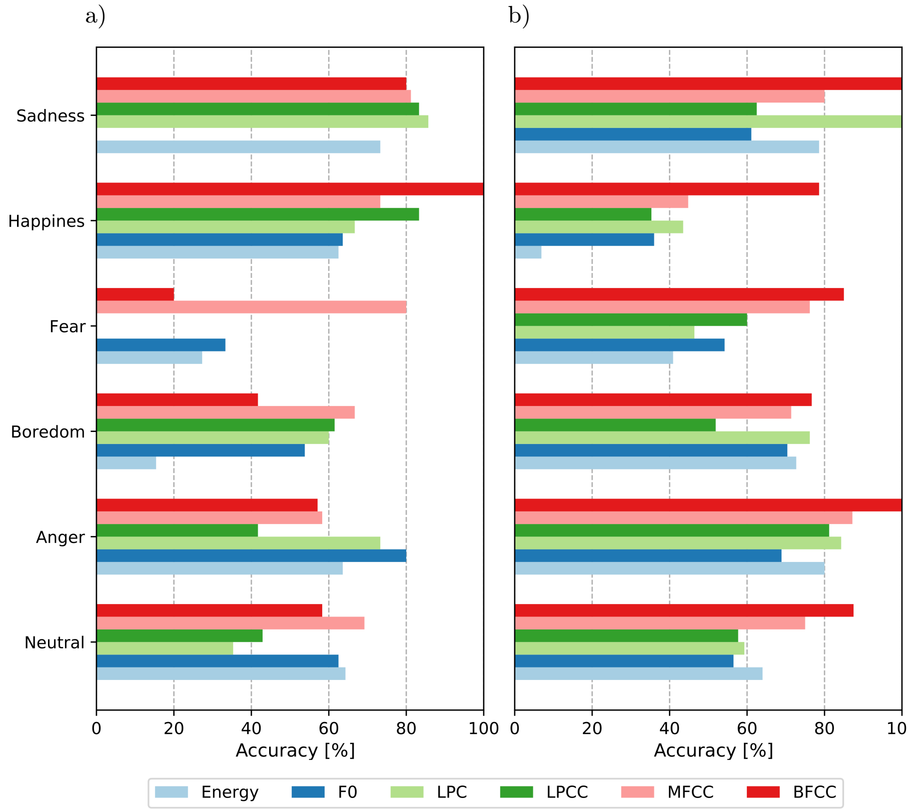

Fig. 3. Classification accuracy using tuned SVM classifier and obtained for individual, reduced feature sets (depicted in Table 2 and Table 3) for dbPL (a) and dbDE (b) datasets.

space. Interestingly, the occurrence of derivatives in the feature space is highly dependent on the source audio data. The best results for each feature subset and both datasets were achieved for SVM classifier. Therefore, we decided to use this classifier in the next experiments.

Having selected the classifier, in the next step, we have performed the hyperparameter optimization of the SVM classifier with RBF, linear, and polynomial kernels using cross-validation. For this purpose we employed GridSearchCV algorithm (Pedregosa *et al.*, 2011). The initial parameters was configured as $C = \{10^{-5}, 10^{-4}, ..., 1, ..., 10^5, 10^6\}$, $\gamma = \{10^{-5}, ..., 1, ..., 10^6, 1/K\}$. In the result, the best configurations were obtained for RBF kernel with $\gamma = 1/K$ and $C = 1$ for dbPL and RBF kernel with $\gamma = 0.01$ and $C = 10$ for the dbDE datasets. After using the tuned classifier for each dataset, the classification results are presented in Fig. 3 where the accuracy for the individual subsets of features can be observed. The final classification accuracy for tuned SVM classifier was equal to 72.5% in case dbPL dataset and 88.27% for dbDE.

4. Discussion

In case dbPL dataset, using the MFCC coefficients, the highest overall correctness was obtained. The highest classification accuracy occurs for *fear*, *boredom*, and *neutral* emotions. However, for *happiness*, the best result was achieved by BFCC coefficients. On the other hand, for the *anger* emotional state, the highest accuracy occurs in subset calculated based on the fundamental frequency and *sadness* emotion was best described by the LPC coefficients. The subset calculated based on BFCC trajectories gave the best results in case of dbDE dataset. The highest classification accuracy was obtained in case of all emotions. Interestingly, the LPC coefficients also ensured the highest classification accuracy for *sadness* emotional state.

Summarizing the results, the most recognizable emotions with the accuracy over 80% were *neutral*, *anger*, *fear* and *sadness* for dbDE dataset. Whereas in the dbPL dataset, emotions such as *fear* and *sadness* were recognized with accuracy above 80%. Additionally, to provide more details on the classification process in the best situations for both datasets, the confuse matrices are depicted in Fig. 4. The distribution of statistics selected in feature sets for both cases is shown in Table 5.

In order to assess the actual discrimination power of attributes optimized separately for both languages, we have performed the classification with exchanged feature spaces. After the experiment for this case, the accuracy for dbPL was equal to 56% while for dbDE

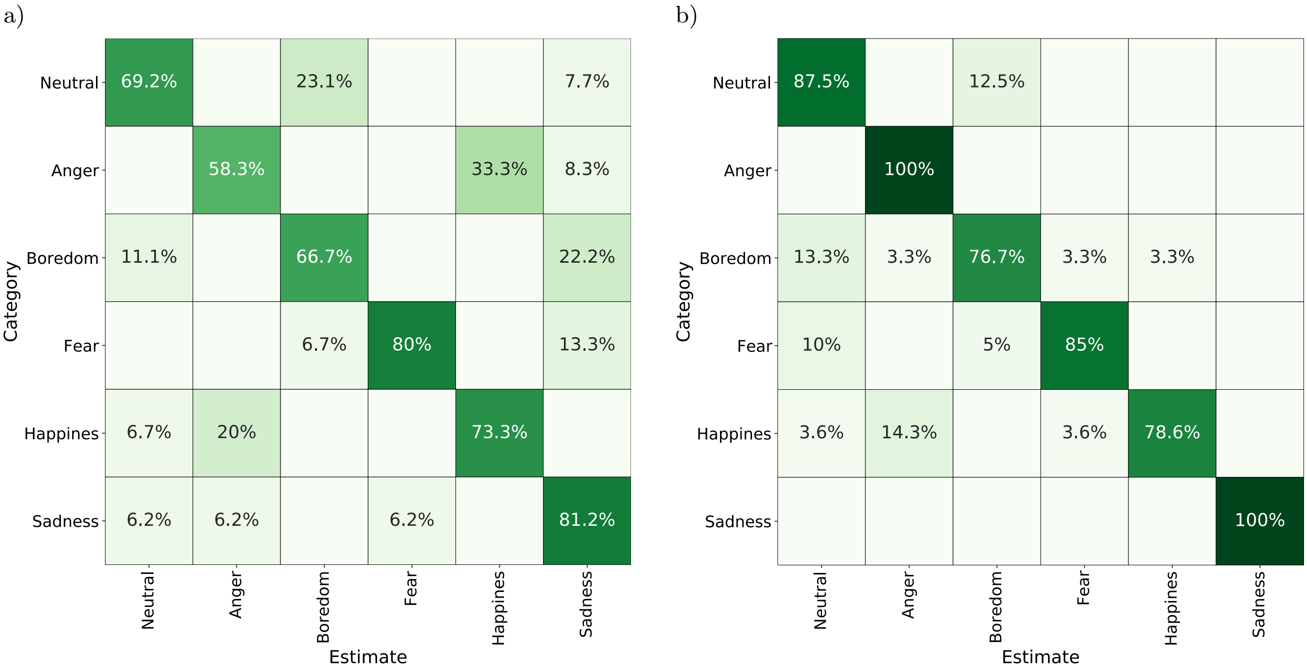

Fig. 4. Confusion matrices for classification using tuned SVM classifier and reduced feature subsets (marked rows in Table 2 and Table 3) for dbPL (a) and dbDE (b) datasets.

Table 5. The number of instances for statistics computed from feature contours in the best cases.

Feature set	MIN	MAX	MEAN	STD	Q1	ME	Q3	RNG	IQR	QD	AD	MAD	VAR
MFCC5 + Δ	4	5	7	9	9	4	9	7	9	9	9	9	9
BFCC20 + ΔΔ	11	17	20	15	20	19	19	13	12	12	15	15	16

was 72%. In both cases, there was a decrease in accuracy equal to 16.5% and 16.3%, respectively. In the case of the dbPL dataset, on average, the most errors occurred as a result of confusing *sadness* emotion with other emotions. For the dbDE dataset, errors most frequently appear while confusing with *neutral* state. In this case, also it occurred a situation where the *anger* and *sadness* emotional states were recognized with perfect accuracy.

5. Conclusions

Emotional state detected from speech signal plays an important role in human-machine interaction systems as an essential component improving the functionality of voice dialogue systems. We described the impact of selected speech signal features and classifications methods on the quality of recognition of the speaker's emotional states. We have shown how individual audio features affect the recognition of specific emotion, which are the most and the least recognizable. Due to the close connection with the voice source, we were expecting at the beginning, a more discriminatory power of attributes related to the fundamental frequency. The result of the analysis turned out to be that the energy distribution in various bands led to better classification results. The less influence of fundamental frequency on the final classification accuracy may be caused by multiple factors related to speakers. The introduction of vocal tract normalization stage along with sophisticated mid- and high-level features based on properties of vocal tract may improve the overall discriminatory power. The differences in the results between the used datasets show how much the expression of specific emotions can differ in two different languages. The dissimilarities between classification results for both cases may arise from the fact that selected languages came from two different language groups where prosodic attributes in expressive speech vary significantly. Another reason may be a quite diverse group of speakers in both databases and a higher number of examples in the dbDE dataset. The obtained results were compared for each used feature subset in terms of a different impact on the classification of specific emotions depending on the speaker's language. In our future work, we plan to build a hybrid approach combining acoustic and linguistic features along with various ensemble learning schemes.

References

1. Anagnostopoulos C.N., Iliou T., Giannoukos I. (2015), Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial*

Intelligence Review, **43**: 155–177, doi: 10.1007/s10462-012-9368-5.

2. Boersma P. (1993), Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *Proceedings of the Institute of Phonetic Sciences*, **17**(1193): 97–110.
3. Boersma P., Weenink D. (2001), Praat, a system for doing phonetics by computer, *Glot International*, **5**(9/10): 341–345.
4. Breiman L. (2001), Random forests, *Machine Learning*, **45**(1): 5–32, doi: 10.1023/A:1010933404324.
5. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. (2005), A database of German emotional speech, *9th European Conference on Speech Communication and Technology*, **5**: 1517–1520.
6. Chang C.-C., Lin C.-J. (2011), LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, **2**: 27:1–27:27, doi: 10.1145/1961189.1961199.
7. Davis S., Mermelstein P. (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**(4): 357–366, doi: 10.1109/TASSP.1980.1163420
8. Eyben F. (2016), *Real-time speech and music classification by large audio feature space extraction*, Springer, Cham, doi: 10.1007/978-3-319-27299-3.
9. Feraru S.M., Zbancioc M.D. (2013), Emotion recognition in Romanian language using LPC features, [in:] *2013 E-Health and Bioengineering Conference (EHB)*, pp. 1–4, doi: 10.1109/EHB.2013.6707314.
10. Hao M., Tianhao Y., Fei Y. (2019), The SVM based on SMO optimization for speech emotion recognition, [in:] *2019 Chinese Control Conference (CCC)*, pp. 7884–7888, doi: 10.23919/ChiCC.2019.8866463.
11. Kathiresan T., Dellwo V. (2019), Cepstral derivatives in MFCCs for emotion recognition, [in:] *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, pp. 56–60, doi: 10.1109/SIPROCESS.2019.8868573.
12. Kuan T.-W., Tsai A.-C., Sung P.-H., Wang J.-F., Kuo H.-S. (2016), A robust BFCC feature extraction for ASR system, *Artificial Intelligence Research*, **5**(2): 14–23, doi: 10.5430/air.v5n2p14.
13. Lee K.H., Kyun Choi H., Jang B.T., Kim D.H. (2019), A study on speech emotion recognition using a deep neural network, [in:] *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1162–1165, doi: 10.1109/ICTC46691.2019.8939830.
14. Markel J.D., Gray A.H.J. (1976), *Linear Prediction of Speech*, New York: Springer-Verlag.
15. Meng H., Yan T., Yuan F., Wei H. (2019), Speech emotion recognition from 3D log-mel spectrograms with deep learning network, *IEEE Access*, **7**: 125868–125881, doi: 10.1109/ACCESS.2019.2938007.
16. Mitrovic D., Zeppelzauer M., Breiteneder C. (2010), Features for content-based audio retrieval, *Advances in Computers*, **78**: 71–150, doi: 10.1016/S0065-2458(10)78003-7.
17. Pedregosa F. *et al.* (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**: 2825–2830, doi: 10.5555/1953048.2078195.
18. Rajak R., Mall R. (2019), Emotion recognition from audio, dimensional and discrete categorization using CNNs, [in:] *TENCON 2019 – 2019 IEEE Region 10 Conference (TENCON)*, pp. 301–305, doi: 10.1109/TENCON.2019.8929459.
19. Rao K.S., Reddy V.R., Maity S. (2015), *Language Identification Using Spectral and Prosodic Features*, Springer Publishing Company, Incorporated.
20. Slot K., Cichosz J., Bronakowski L. (2009), Application of voiced-speech variability descriptors to emotion recognition, [in:] *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–5, doi: 10.1109/CISDA.2009.5356537
21. Swain M., Routray A., Kabisatpathy P. (2018), Databases, features and classifiers for speech emotion recognition: a review, *International Journal of Speech Technology*, **21**: 93–120, doi: 10.1007/s10772-018-9491-z.
22. Ververidis D., Kotropoulos C. (2006), Emotional speech recognition: Resources, features, and methods, *Speech Communication*, **48**: 1162–1181, doi: 10.1016/j.specom.2006.04.003
23. Zhang H. (2004), The optimality of naive bayes, [in:] *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, FLAIRS 2004.
24. Zhu C., Ahmad W. (2019), Emotion recognition from speech to improve human-robot interaction, [in:] *2019 IEEE International Conference on Dependable, Autonomic and Secure Computing*, pp. 370–375, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00076.