**Dorota KAMIŃSKA, Tomasz SAPIŃSKI**

Lodz University of Technology, Institute of Mechatronics and Information Systems

doi:10.15199/48.2017.06.24

# Polish Emotional Speech Recognition Based on the Committee of Classifiers

***Abstract**. This article presents the novel method for emotion recognition from polish speech. We compared two different databases: spontaneous and acted out speech. For the purpose of this research we gathered a set of audio samples with emotional information, which serve as input database. Multiple Classifier Systems were used for classification, with commonly used speech descriptors and different groups of perceptual coefficients as features extracted from audio samples..*

***Streszczenie.** Niniejsza praca dotyczy rozpoznawania stanów emocjonalnych na podstawie głosu. W artykule porównaliśmy mowę spontaniczną z mową odegraną. Na potrzeby zrealizowanych badań zgromadzone zostały emocjonalne nagrania audio, stanowiące kompleksową bazę wejściową. Przedstawiamy nowatorski sposób klasyfikacji emocji wykorzystujący komitety klasyfikujące, stosując do opisu emocji powszechnie używane deskryptory sygnału mowy oraz percepcyjne współczynniki hybrydowe. **(Rozpoznawanie emocji na podstawie sygnału mowy przy użyciu komitetów klasyfikujących)***

**Keywords**: emotion recognition, speech, Committee of Classifiers, k-NN.
**Słowa kluczowe**: rozpoznawanie emocji, mowa, komitet klasyfikujący, k-NN.

## Introduction

Emotions are a carrier of information regarding feelings of an individual and one's expected feedback. Understanding them enhances interaction. Although computers are now a part of human life, the relation between a human and a machine is not natural. Knowledge of the emotional state of the user would allow the machine to adapt better and generally improve cooperation between them.

Emotion recognition methods utilize various input types i.e. facial expressions [1], speech, gesture and body language [2], physical signals such as electrocardiogram (ECG), electromyography (EMG), electrodermal activity, skin temperature, galvanic resistance, blood volume pulse (BVP), and respiration [3]. Speech is one of the most accessible form the above mentioned signals and because of this most emotion recognition related research focuses on human voice, which became a relevant trend in modern studies. However, satisfactory efficiency has not been achieved yet.

This article presents a novel method of emotional speech recognition based on k-NN utilizing multiple classifies. We compared two alternative approaches using equal and weighted voting. The research has been conducted on two different sets of Polish emotional speech: acted out and spontaneous. A pool of descriptors, commonly utilized for emotional speech recognition, expanded with sets of various perceptual coefficients, has been used in as input feature vectors.

The outline of the paper is as follows. Section II presents a brief review of selected studies conducted in this area. Next sections describe proposed research methodology: present the theory of emotion, database of emotional voice, speech signal descriptors and outlines adopted strategy for emotion recognition. Section VII presents obtained results. Finally, section VIII gives the conclusion and future directions of this research.

## Related works

Emotion recognition from speech is a pattern recognition problem. Therefore, standard pattern recognition methodology, which involves feature extraction and classification, is used to solve the task. 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 are utilized. Therefore, fundamental frequency, formants, energy, mel-frequency cepstral coefficients (MFCC) or linear prediction coefficients (LPC) are widely explored [4]. Nevertheless the search of new speech features is ongoing [5] [6] [7]. However, a vector of too many features does not necessarily lead to a more accurate prediction [8]. Therefore methods of balancing a numerous features vector and emotion classification accuracy are studied [9] [10].

Recognition of feature vectors is generally performed using well known algorithms, starting from vector classification methods, such as Support Vector Machines [11] [12], various types of Neural Networks [13] [14], different types of the k-NN algorithm [15] [16] or using hidden Markov model (HMM) and its variations [17]. Some scientists create multimodal or hierarchical classifiers by combining existing methods in order to improve recognition results [18].

Most of the research is based on Berlin emotional database, which is a standard for emotion detection [19]. It contains speech samples of six emotions: anger, fear, happiness, sadness, disgust, boredom and neutral state. Other public licensed corpora also focus on this spectrum of emotions. The usage of such database with well-defined emotional content leads to high performance of classification algorithm. However, emotion classification system designed for real-world application must be able to interpret the emotional content of an utterance during spontaneous dialogue, which involves a complex range of mixed emotional expressions [20]. As was shown in [21] natural human speech may be inherently ambiguous. Thus, a conventional designed system for emotion recognition based on acted speech may not be able to handle the natural variability of spontaneous speech..

Recent studies have shown that an efficient database should consists of natural emotions, which contain valuable material reflecting the speaker's spontaneous reactions. However, creation of such database proves to be a difficult task. Typically it is recorded provoking an appropriate response e.g. controlling a computer game scenario or recording speakers in natural situations, such as interaction with a robot. These recordings usually contain merely narrow spectrum of emotional states, often of poor quality, containing background noise, etc. Creation of a spontaneous speech database involves samples labelling, which is usually performed by volunteers - listeners. Evaluation of opposing emotions does not prove to be a difficult task [22]. However, assessment of similar states, appears to be dependent on subjective feelings of the listener. Therefore, it should take place in the large group of of decision makers, which makes this process time-consuming.

**Theory of Robert Plutchik**

This paper refers to the theory of Robert Plutchik [20], an American psychologist. He suggested, that eight basic congenital emotions exist, related biologically to the adaptation for survival. From the merger of primary emotions more complex emotions are formed. Figure 1 shows the relationship between emotions according to Plutchik’s theory. Primary emotions are presented on the base of the cone as pairs of opposites i.e. sadness is vis-a-vis to joy, anger is opposite to fear etc. The cone’s vertical dimension represents intensity i.e. anger, which starts from annoyance, can intensify to rage. From the merger of primary emotions more complex states are formed i.e. a combination of anticipation and joy results in optimism.

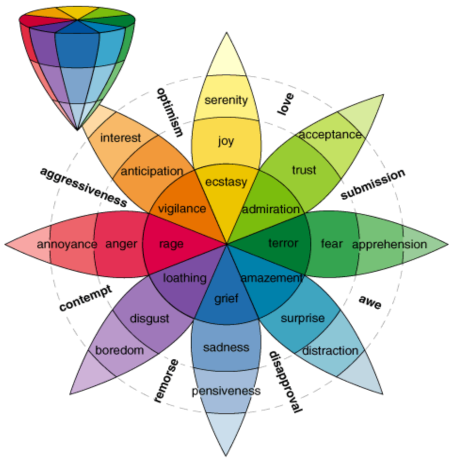


Fig.1. Plutchik’s wheel of emotions [20]

**Emotional speech corpora**

For the purpose of this research we gathered audio samples containing basic states from Plutchik's wheel of emotions: joy, sadness, anger, fear, disgust, surprise and anticipation. Additionally a set of neutral speech has been created. This resulted in more than 700 speech samples evenly distributed across all emotions. Recordings have been collected from various Polish TV programs, mostly reality shows, which are characterized by occurrence of spontaneous emotional dialogues. Moreover, samples have been recorded in various forms i.e. in noisy environments or with imperfect sound quality. Therefore the created database can be a good reflection of the real world conditions. All samples were assessed by a group of human evaluators (experts and volunteers) and labeled into the above mentioned classes of emotions. The division process was carried out using video material, thus features such as facial expressions, gestures and semantics facilitated the process of labelling. Only those samples which were most clearly identified have been chosen for further analysis. The whole process of creation Polish emotional database was fully described in [23].

We also use Polish Acted Emotional Speech for comparison. It was 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 uttered five different sentences with five types of emotional load: joy, boredom, fear, anger, sadness and neutral state. Recordings were taken in the aula of the Polish National Film Television and Theater School in Lodz, using a condenser microphone. The audio files are available in the PCM WAVE format file with a 44100 Hz sampling rate [24].

**Speech signal parameters**

Basing on speech production model prosodic and vocal tract descriptors have been used as components of the features vector. Prosody is one of the most important communicative channels and plays a major role in expressing emotions in speech. The standard approach in emotion recognition is based on prosodic information processing (fundamental frequency and signal energy) and articulation (formants).

Fundamental frequency (pitch), which defines the speech source, describes tonal and rhythmic properties of speech. From a variety of methods of determining fundamental frequency, autocorrelation algorithm has been used for the purpose of this research. Objective assessment of F0 behavior might prove to be difficult basing solely on its contour. Therefore, rather than using the pitch value itself, it is commonly accepted to utilize global statistical features of the pitch contour over an entire utterance.

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 have increased energy levels in comparison to other emotional states.

Formant frequencies are the frequencies, at which there are local maxima of the speech signal spectrum envelope. They are the properties of vocal tract. Basing on them, it is possible to determine who the speaker is and what and how he is speaking. In practical applications 3 to 5 formants are used. In this paper 3 formant frequencies were estimated.

LPC is a time domain technique that models a signal as a linear combination of weighted delayed values. For speech signal linear predictive coding is used to calculate a set of coefficients, which approximate the vocal tract. For the purpose of this project 13 LPC coefficients were used.

Perceptual approach is based on frequency conversion, corresponding to subjective reception of the human hearing system which does not follow a linear scale. For the purpose of this research, the perceptual Mel and Bark scales are used. In this paper Mel Frequency Cepstral Coefficients (MFCC), Bark Frequency Cepstral Coefficients (BFCC), Perceptual Linear Prediction Coefficients (PLP) and Revised Perceptual Linear Prediction Coefficients (RPLP), were taken into consideration. Perceptual approach in emotion recognition is presented in [25].

We obtained a large set of various features and in order to distinguish the most discriminative ones, the whole set was subjected to the process of selection. Taking into account the type of classification method with the division into sub-features, the selection was conducted on specific subsets, and to compared with the whole set. The Sequential Backward Selection (SFS) method was used with k-NN classifier as a criterion function. Reduction of dimensionality was achieved: from 448 attributes to 57 (whole set selection) and 98 (subsets selection) in case of the acted speech database, from 473 attributes to 88 (whole set selection) and 160 (subsets selection) in case of the spontaneous speech database. Example results of experiments obtained for spontaneous speech are presented in Table 1.

Table 1. Most discriminative features obtained for spontaneous speech corpora using SFS selection

| Features | Whole set selection | Subsets selection |
|---|---|---|
| Fundamental frequency | mean, median, kurtosis, skewness, variation rate | mean, median, max, range, upper quartile, lower quartile, interquartile range, kurtosis, skewness, variation rate |
| Formants | mean:f1;f3; max:f3; | mean:f1;f2; max:f2; median:f1;f2,min:f3; std f3 |
| Energy | max, min, median, std, | max, min, median, std, range |
| LPC | LPC 6,11 - mean; LPC 6: median; LPC 4,9 - std; LPC 9,10: min | LPC 2,3,4,7,9 - mean; LPC 2: median |
| BFCC | BFCC:9,10 - mean; BFCC: 2,5,7 - median; BFCC:9,11 - std; BFCC: 9 - max; | BFCC:1,4,5,6,7, 8, 9,11,12 - mean; BFCC: 1,2,3,4,5,8 - median; BFCC:1 - std; BFCC: 1,4,12- max; BFCC: 2 - min |
| MFCC | MFCC: 1 - std; MFCC: 2 - median; MFCC: 1, 9 - min | MFCC: 1,2,3,4,6, 9,10 - mean; MFCC: 1,7,11,12 - std; MFCC: 3,4,7 - median; MFCC:1,2,4,5 - max MFCC: 1,3,5,8,12- min |
| HFCC | HFCC: 1, 2 - mean; HFCC: 5 - std; HFCC:1,3,4,6,9 median; HFCC: 3,4,5,7 -max; | HFCC: 1,3,4,6,7, 9,10 - mean; HFCC:2,3,11 - std; HFCC:1,2,3,4,6,8 median HFCC: 1,2,3,5,7 -max; HFCC: 8 - min |
| PLP | PLP: 4,6,9 - mean; PLP: 7, 8, 9 -median; PLP: 4, 9-std; | PLP: 4,5,9,13 - mean; PLP:8,10,11,12,13 - median; PLP: 3,7,12- std; PLP: 6,12,13 - max; PLP: 1,2,4,6,9,10,11,12 - min |
| RPLP | RPLP: 2,4,5,9,12 - mean; RPLP:4,6,9,12 - median; RPLP 1,2,3,4,5,11,13 - std; RPLP: 1,2,3 - max; | RPLP: 1,2,5,14 - mean;RPLP:1,2,3,5,6, 7,9,14 - median; RPLP:1,2,4,5,10,11,14 - std; RPLP: 1,4,6, - max;RPLP:1,2,3,11, 13 - min |
| RASTA PLP | RASTA: 5,7 - mean; RASTA: 9,10- median; RASTA: 2,6,7,8,9,10 - std; RASTA: 6,8,9,10 - max; RASTA :1,2,4,7 - min | RASTA: 4,9 - mean; RASTA: 2,3,5,8 - median; RASTA: 1,3,5,10 - std; RASTA: 6,9 - max; RASTA:1,2,3,4,7 - min |

**Strategy for emotion recognition**

Committees (Multiple Classifier Systems - MCS) are based on divide and conquer principle, decomposing a complex problem into several smaller [26].
This solution consists of many simple (easy to build) models (nodes) with relatively low efficiency which final results are combined (e.g. by voting). Individual modules are created independently or sequentially. Performance of such a system is connected with its hierarchical structure rather than the computing power of the individual components. An additional advantage is the extensibility and simplicity of the system.

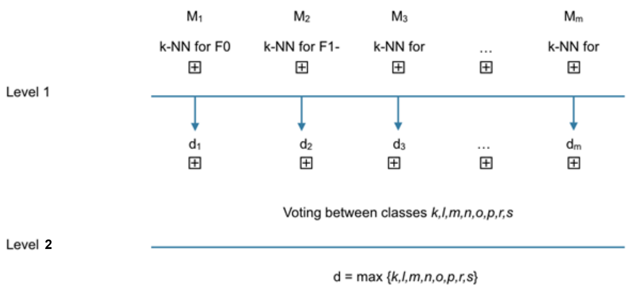


Fig.2. Proposed algorithm for emotion recognition using committee of classifiers

Proposed algorithm (Fig. 2) works as follows. Feature vectors describing objects from the training set and the unknown object were divided into sub-vectors of particular group of features (Tab. 1). Basing on that assumption we created m separate sub-vectors (i.e. vector with MFCC coefficients). M different models (*k*-NN) produce m decisions basing on different set of features on the first stage of the algorithm. The final decision is obtained on the second level of the algorithm using voting basing on m different decisions (where *k,l,m,n,o,p,r,s* represents numbers of votes for particular class).

If there is no possibility to emerge the winning class basing on equal voting, we use the Plutchick's wheel to make the final decision. If both winning class lie on the wheel next to each other, it is likely that the boundaries between them are so ambiguous that it is not possible to assign one specific class. Such an utterance is classified as ambiguous and labelled with two emotions classes. The polar opposite of winning classes disqualify (according Plutchik's theory) the possibility of occurrence two of states at the same time. In this case the winning class is chosen basing on the performance of each subset (Fig. 3).

In the basic algorithm the final decision is made using equal voting. This method does not demand additional calculations, just votes of individual models, rendering this process simple and effective [27]. A decision is made collectively $Z = T_1, T_2 \ldots T_k$ using following equations:

(1) $$r_i = \sum_{j=1}^{k} d_{ji}$$

(2) $$Z = \arg\max_{i=1}^{l}[r_i]$$

where: $m$ – number of classifiers, $l$ – number of different classes, $d_{ji}$ – decision of $j$ classifier for $i$ class.

Unequal impact of particular descriptors on the recognition provides the basis for replacing equal voting by weighted in the second approach. In this type of voting votes are not counted equally. For each model we determined different weights $w_1, w_2, \ldots w_m$, which allow to prioritise more precise models. In this case the final decision is made using following equations:

(3) $$r_i = \sum_{j=1}^{k} w_j d_{ji}$$

(4) $$Z = \arg\max_{i=1}^{l}[r_i]$$

This approach requires the skill of assessing (or at least comparing) all models. In this study weights were selected experimentally, based on the error of individual classifiers (Fig.3).

**Achieved results**

Presented method of classification was verified using two polish emotional speech corpora. The average recognition rate of emotions evaluated by a group of experts amounts to 82.6% in case of spontaneous speech and 72% in case of acted out speech. These results represent a benchmark for recognition algorithms. In case of spontaneous speech it is not possible to ensure the recurrence of registered materials, we cannot perform speaker dependent research. Therefore, in this work we focus only on the speaker independent analysis. For this purpose the whole set was divided into the test set (33% of all the samples) and the training set (66%).

In the first step we compared the recognition rates of our method with single k-NN algorithm. The study was conducted on the entire pool of features before the selection process. The k value for k-NN was chosen experimentally and the best results were achieved with k=3 for both datasets. Than we used SFS selection and basing on obtained dominative features (Tab.1) we compared both algorithms. Best results were obtained for k=3 for acted out speech and k=5 for spontaneous speech. In the case of proposed algorithm the k value was selected experimentally for each subset of features and presents in Table 2. Results of classification are presented in Table 3.

Table 2. K value selected experimentally for each subset of features for proposed algorithm PA

| Group of features | k value | k value SFS | k value | k value SFS |
|---|---|---|---|---|
| | spontaneous speech | | acted out speech | |
| F0 | 3 | 7 | 7 | 3 |
| Formants | 1 | 1 | 5 | 1 |
| Energy | 7 | 3 | 7 | 1 |
| LPC | 1 | 1 | 3 | 1 |
| BFCC | 3 | 5 | 1 | 3 |
| MFCC | 9 | 1 | 9 | 3 |
| HFCC | 5 | 5 | 1 | 3 |
| PLP | 1 | 1 | 3 | 1 |
| RPLP | 1 | 9 | 1 | 1 |
| RASTA | 7 | 1 | 9 | 9 |

Table 3. Comparison of average recognition rates of k-NN and proposed algorithm (PA) before and after selection

| Method | Accuracy performance for spontaneous speech database[%] | Accuracy performance for acted out speech database [%] |
|---|---|---|
| k-NN | 78,9% | 60,8% |
| k-NN SFS | 78,9% | 65,8% |
| PA | 79,3% | 65,8% |
| PA SFS | **81,6%** | **70,9%** |

Table 4. Confusion matrix of emotion recognition using k-NN/proposed algorithm for spontaneous speech: A - anger, An - anticipation, D - disgust, F - fear, J - joy, Sd - sadness, Su - surprise

| | A | An | D | F | J | Sd | Su |
|---|---|---|---|---|---|---|---|
| A | **30/30** | 1/2 | 0/0 | 0/0 | 1/1 | 5/4 | 0/0 |
| An | 0/0 | **28/27** | 0/0 | 0/2 | 0/0 | 0/0 | 1/0 |
| D | 2/0 | 0/0 | **23/27** | 1/1 | 1/1 | 2/0 | 1/1 |
| F | 2/1 | 1/1 | 0/0 | **30/37** | 5/2 | 2/0 | 3/2 |
| J | 3/3 | 1/1 | 0/0 | 5/5 | **44/46** | 1/0 | 1/0 |
| Sd | 6/2 | 1/3 | 0/0 | 0/0 | 1/2 | **30/30** | 0/1 |
| Su | 0/2 | 5/6 | 0/0 | 1/2 | 1/2 | 1/1 | **21/16** |

Table 5. Confusion matrix of emotion recognition using k-NN/proposed algorithm for acted out speech: A - anger, B - boredom, F - fear, J - joy, Sd - sadness, N - neutral state

| | **A** | **B** | **F** | **J** | **Sd** | **N** |
|---|---|---|---|---|---|---|
| **A** | **7/11** | 2/0 | 0/0 | 3/1 | 0/0 | 1/1 |
| **B** | 1/0 | **8/6** | 0/0 | 0/0 | 2/3 | 2/4 |
| **F** | 1/2 | 1/3 | **7/6** | 0/0 | 3/2 | 1/0 |
| **J** | 1/1 | 0/0 | 2/1 | **11/12** | 0/0 | 0/0 |
| **Sd** | 1/0 | 2/2 | 0/0 | 0/0 | **10/11** | 1/1 |
| **N** | 0/0 | 2/2 | 0/0 | 0/0 | 1/0 | **9/10** |

According to Table 3 we can observe a slight improvement of recognition rate using proposed algorithm in comparison to ordinary k-NN method for both datasets. Detailed results for both algorithm are presented by confusion matrix (Tab.4 and Tab.5).

The analysis of confusion matrix shown that the main mistakes are made in the case of anger-sadness, boredom-sadness, anticipation-surprise for spontaneous database and anger-joy, boredom-sadness-neutral state and fear-sadness for acted out database.

Since each of the subsets of features has different impact on recognition, equal voting in proposed algorithm was replaced by weighted, thereby increasing the influence of more discriminant sets of features. For this purpose we selected appropriate weight $w_i$ for individual classifier. Those weights were calculated based on the error of particular model $err_i$ on training test using 10-fold cross validation according to following equation.

$$w_i = 1 - err_i \tag{5}$$

$$w_i = \frac{1}{err_i} \tag{6}$$

$$w_i = \left(\frac{1}{err_i}\right)^2 \tag{7}$$

The results are compared in Table 6. Figure 3 presents errors values obtained for each subset for both corpora.

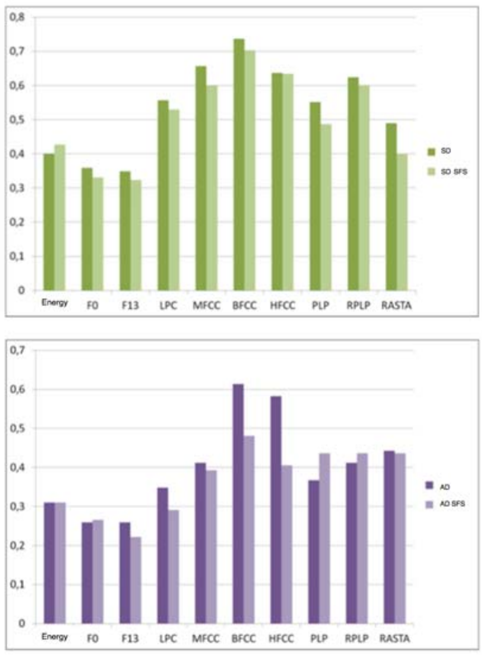


Fig.3. Errors value for each subset: top - Spontaneous Database SB, bottom Acted out Database AD

Table 6. Comparison of accuracy performance of proposed algorithm using equal and weighted voting

| Voting type | AP [%] for spontaneous speech | AP [%] for acted out speech |
|---|---|---|
| Equal | 81,6 | 70,9 |
| Weighted Eq. 5 | 84,3 | 74,7 |
| Weighted Eq. 6 | **83,9** | **75,9** |
| Weighted Eq. 7 | **84,7** | 67 |

Analyzing above results one can observe improved results after using weighted voting. For spontaneous speech the best results are obtained using equation 7, for acted out speech equation 6.

Moreover, one can notice significantly lower results in case of acted speech. This is the result of the different contents of the two databases. In the case of acted speech corpora, the relative number of speakers, in comparison to the sample count, could affect the performance of the classifiers.

## Conclusions

The main assumption of these research was to develop a system allowing automatic emotional state recognition based on natural speech. To achieve this we created a polish spontaneous emotions database consisting of over 700 samples divided into sets representing primary emotional states. Description of the problem is based on features commonly used in this type of research, these speech descriptors combined with hybrid perceptual coefficients were never before applied in this area. As the research has shown, these features proved to be highly discriminative what justifies their application. During the classification we compared k-NN algorithm with an original approach based on a set of classifiers (committee) ensuring better recognition results. The analysis proved the initial assumptions.

***Authors**: dr inż. Dorota Kamińska, Lodz University of Technology, IInstitute of Mechatronics and Information Systems, Stefanowskiego 18/22 Str., 90-924 Lodz, E-mail: dorota.kaminska@p.lodz.pl; mgr inż. Tomasz Sapiński, Lodz University of Technology, IInstitute of Mechatronics and Information Systems, Stefanowskiego 18/22 Str., 90-924 Lodz, E-mail: tomasz.sapinski@p.lodz.p*

LITERATURE

[1] Metallinou A., Katsamanis, Narayanan S., A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, (2012), 2401-2404
[2] Gunes H., Piccardi M., Bi-modal emotion recognition from expressive face and body gestures, J. *Netw. Comput. Applicat*, 30(4), (2007), 1334-1345
[3] Garay N., Cearreta I., Lpez J.M., Fajardo I., Assistive technology and affective mediation, *Assistive Technol*, 2(1), (2006), 55 – 83
[4] Deng J., Han W., Schuller B., Confidence Measures for Speech Emotion Recognition: A Start, *Proceedings of Speech Communication Symposium (2012)*
[5] Ntalampiras S., Fakotakis N., Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition, IEEE Transaction on Affective Computing, 3 (2012), 116 – 124
[6] Christina I.J., Milton A., Analysis of All Pole Model to Recognise Emotion From Speech Signal, International Conference on Computing, Electronics and Electrical Technologies, (2012), 723 – 728
[7] Ślot K., Rozpoznawanie biometryczne - Nowe metody ilościowej reprezentacji obiektów, (2010), Warszawa
[8] Fewzee P., Karray F., Dimensionality Reduction for Emotional Speech Recognition, 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, (2012), 532 – 537
[9] Ivanov A., Riccardi G., Kolmogrov-Smirnov test for feature selection in emotion recognition from speech, 38th International Conference on Acoustics, Speech, and Signal Processing, (2012), Canada, 5125 – 5128
[10] Han W., Zhang Z., Deng J., Wollmer M., Weninger F., Schuller B., Towards Distributed Recognition of Emotion From Speech, Proceedings of the 5th International Symposium on Communications, Control and Signal Processing, (2012) Rome, Italy
[11] Yup S., and Yoo C., Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification, IEEE Transactions on audio, speech and language processing 20(2), (2012)
[12] Yang N., Muraleedharan R., Kohl J., Demirkoly I., Heinzelman W., Sturge-Apple M., "Speech-based emotion classification using multi class SVM with hybridkernel and thresholding fusion, The Fourth IEEE Workshop on Spoken Language Technology (SLT) (2012), Miami
[13] Bojani M., Crnojevi W., Deli V., Application of Neural Networks in Emotional Speech Recognition", 11th Symposium on Neural Network Applications in Electrical Engineering, (2012) 223 - 226, Belgrade
[14] Powroźnik P., Czerwiński D., Effectiveness comparison on an artificial neural networks to identify Polish emotional speech, Przegląd Elektrotechniczny, 7, (2016), 45-48
[15] Attar Y., Dumouchel P., Emotion recognition from speech: WOC-NN and class-interaction", 11th International Conference on Information Science, Signal Processing and their Applications, (2012), 126 – 131
[16] Zbancioc M., Feraru S. M., Emotion Recognition of the SROL Romanian Database using Fuzzy KNN Algorithm", 10th International Symposium on Electronics and Telecommunications Proc., (2012)Timisoara
[17] Metallinou A., Katsamanis A., Narayanan S., A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs, 38th International Conference on Acoustics, Speech, and Signal ProcessinG, (2012), Canada, 2401 – 2404
[18] Vasuki P.,Aravindan C., Improving Emotion Recognition From Speech Using Sensor Fusion Techniques, TENCON, (2012), Cebu, Philippines
[19] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B., A database of german emotional speech, InterSpeech, (2005), 1517 – 1520
[20] Plutchik R., Nature of emotions, American Scientist, 89,2001
[21] Mower E., Metallinou A., Lee C., Kazemzadeh A., Busso C., Lee S., Narayanan S., Interpreting ambiguous emotional expressions, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (2008), 1-8
[22] Kamińska D., Pelikant A., Recognition of Human Emotion from a Speech Signal Based on Plutchik's Model, International Journal of Electronics and Telecommunications 58 (2), (2012),165-170
[23] Kamińska D., Sapiński T., Pelikant A, Polish Emotional Natural Speech Database, Signal Processing Symposium, (2015), Debe, 63-69
[24] Database of Polish Emotional Speech, available: http://www.eletel.p.lodz.pl/bronakowski/med_catalog/
[25] Kamińska D., Sapiński T.,Pelikant A., Comparison of Perceptual Features Efficiency for Automatic Identification of Emotional States from Speech, HSI.2013
[26] Cichosz P. Systemy uczące się Wydawnictwa Naukowo-Techniczne, 2007
[27] Podolak I. Klasyfikator Hierarchiczny z nakładającymi się grupami klas. Wydawnictwo Uniwersytetu Jagiellońskiego, 2012