Kohonen network as a classifier of Polish emotional speech

Paweł Powroźnik[1,*]

[1]Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Institute of Computer Science, Nadbystrzycka 38D, 20-618 Lublin, Poland

Abstract. The power of speech is a main tool in human communication. There are a lot of factors as age, emotions, gender, pitch of the voice which can influence features of speech. Obviously, information conveyed by voice intonation has more than only textual meaning. The same sentence pronounced in two different ways can have two completely different meanings. This paper describes Kohonen networks as a classifier of Polish emotional speech. The usage of Discrete Wavelet Transform (DWT) as well as an innovative approach to scaleogram processing is also presented in this article. Mexican Hat Wavelet and the Haar Wavelet were used in researches. All simulations were carried out in MatLab 2016 with Neural Network Toolbar. During whole research more than 9000 simulation have been done. Three different speech databases were used in conducted researches. One of them was prepared by professional actors – four women and four men, and contains 240 wav files. Two others are results of researchers works. The structures of used Kohonen networks depend on speech signal decomposition's level and scaleogram division. During conducted researches the following emotional states were considered: anger, joy, sadness, boredom, fear and neutral state. Achieved results were between 68% and 80% depends of used wavelet, speech signal and signal decomposition's level.

1 Introduction

Recognition of speaker's emotional state based on speech signal processing is relatively new issue but its significance has been rapidly increasing. One of the reasons of such a direction of changes is burgeoning development of systems based on Brain-Computer Interface, as well as Virtual Reality (VR) environments [1].

There are a lot of factors such as: age, emotions, language or gender of a speaker which may have great influence on features of speech signal [2, 3]. Obviously, that information conveyed by intonation of voice has more than only literal meaning.

The biggest problem in emotional speech recognition systems is the number of emotional states; therefore, developing an application which correctly identifies most emotions is not trivial. So far, in researches the following emotional states are often considered: anger, joy, sadness, fear, boredom and neutral state [4, 5].

In the Figure 1, oscillograms for three emotional states are presented, including: neutral state (top), joy (middle), anger (bottom) for the same expression in the semantic sense. As it can be easily seen, the appearance of a particular emotion changed not only ranging from frequency but also to the shape of oscillogram.

In this article, an innovative system of Polish emotional speech signal processing has been described. The system based on discrete wavelet transform and scaleograms. As a classifier the Kohonen networks have been used.

The whole article was divided in four parts. The first one charactering the discussed matter. The second one described used databases of Polish emotional speech. The third part included the description of signal processing algorithm, research methods, and parameters for the Polish emotional speech and presented obtained results and suggestions for improving the adapted research methods.

2 Analysis of issues

Nowadays, the researches struggle with developing a proper and useful model of emotion. Contemporarily, two most popular models exist in the literature: James-Lange's and Plutchik's models [3, 4, 6]. The first one assumes that behavioural and somatic changes are interpreted as emotion which triggers a reaction. Based on James-Lange's model the following sentence is true: 'I am afraid of murder because I am running' instead of 'I am running because I am afraid of murder' [7]. The second model was proposed in 1960 by Robert Plutchik. He introduced eight basic emotional states which were related to behaviours essential for surviving, that is: joy, fear, trust, sadness, surprise, anger, disgust and anticipation. All other states arose from the basic ones [8].

As it was mentioned, the most complicated issue in Polish emotional speech recognition is the number of emotional states which should be detected. It should

* Corresponding author: p.powroznik@pollub.pl

be emphasised that for an average person it is possible to recognise another person's emotional state only in 60% of all cases [9]. Some methods of emotions' detection in speech signal, based on the Polish language, have been described in the literature. These methods relate to Support Vector Machine (SVM) [10] or k-Nearest Neighbour algorithms [4]. However, all authors conclude that obtained results are not fully satisfactory and used signal processing methods need to be improved. All mentioned researches focused on six most popular emotional states: anger, joy, boredom, fear, neutral state and sadness [4, 10].

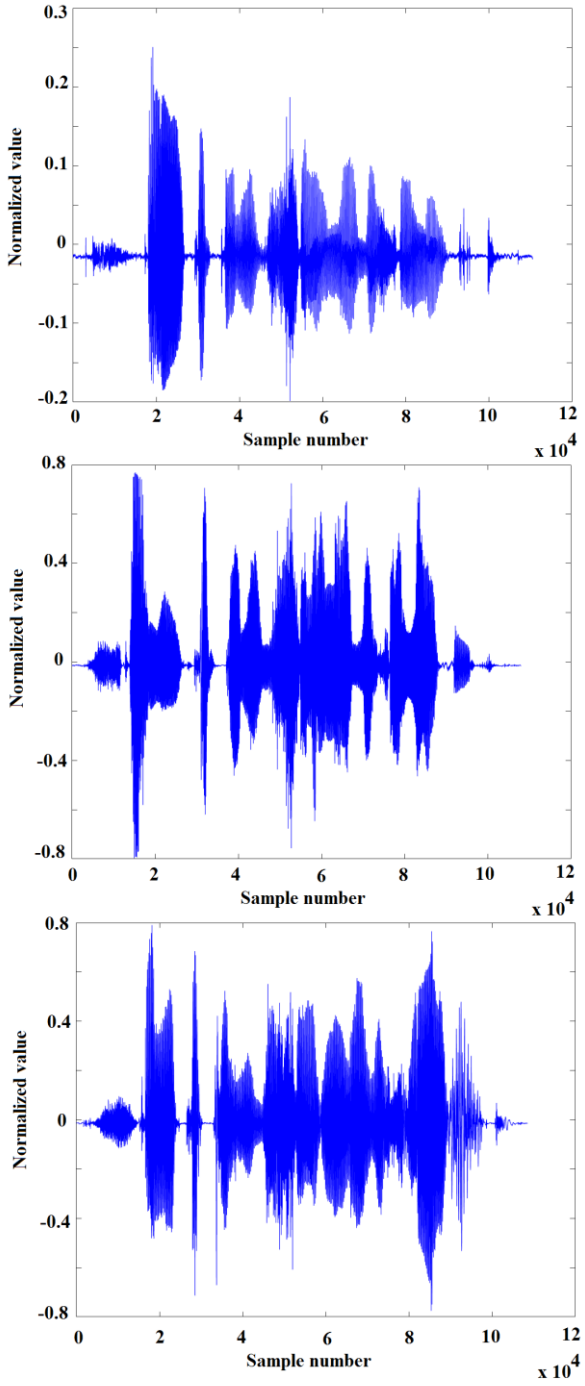

Fig. 1. Comparison of oscillograms of the same sentence in three different emotional states: neutral (top), joy (middle), fear (bottom).

Time-frequency methods are among the most popular speech signal processing tools [11]. They allow to estimate speech signal spectrum in short and finite time based on window functions and overlapping method [11]. A particular role in these cases, has Short Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT) [2, 4, 10, 12]. In this article the usage of second one has been described in detail.

3 Description of used databases

Berlin Database of Emotional Speech (BES) [13] is commonly used in many researches connected with emotional speech signal processing. However, if Polish speech is considered, a researcher would rather use their own databases or the one prepared by Medical Electronics Division of the Lodz University of Technology [14]. The base has been prepared by professional actors: four women and four men, and collected files was recorded in six emotional states. The whole database contains 240 records sampled with 44.1 kHz frequency and the bit rate of 16 bps. This database includes the following statements: 'Johnny went to the hairdresser today', 'They bought a new car today', 'I've stopped shaving from today on', 'His girlfriend is coming here by plane' and 'This lamp is on the desk today'. It was the first database used in conducted researches.

The second one was prepared in Lublin University of Technology and contained the same records as the first one. The abovementioned collection was recorded in acoustic chamber. The research involved people between 20 and 30, who were not involved in acting. The entire database contains 306 records and its structure is presented in the Table 1. Unfortunately, it was not possible to collect the same number of recordings for each emotional state for both sexes.

Table 1. Structure of recordings in second database.

Emotion	Number of recordings (women)	Number of recordings (men)
Anger	14	19
Joy	13	27
Sadness	32	30
Neutral state	28	29
Boredom	31	29
Fear	28	26

The third database was also prepared in Lublin University of Technology but the recordings was collected not in acoustic chamber but in the city environment (lecture class, street, shop). Three women and six men aged 22-31 were involved in this study. They uttered following five sentences: 'I need to talk to you', 'Is this what you really think?', 'You don't understand anything', 'Peter bought a new bike', 'This package is in Krakow already'. Whole database contains 266 records and its structure is shown in the Table 2.

Table 2. Structure of recordings in third database.

Emotion	Number of recordings (women)	Number of recordings (men)
Anger	22	28
Joy	18	20
Sadness	24	23
Neutral state	26	24
Boredom	21	27
Fear	15	18

In case of second and third database, the affiliation of recordings to specific group of emotion was verify by 94 respondents.

4 Conducted researches

During conducted researches, following emotional states were considered: anger, joy, sadness, boredom, fear and neutral state. All simulations have been done in MatLab 2016 with Neural Network Toolbox. The number of all experiments exceeded 9000.

4.1 Wavelet Transformation

Continuous Wavelet Transform (CWT) was developed by Jean Morlet and Alex Grossman. One dimensional signal is expressed by following form [15]:

$$x(t)\epsilon L^2(\mathcal{R}) \tag{1}$$

CWT has the following form [16]:

$$w(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{2}$$

where: * – indicates Conjugate complex function, *a* (*a*>0) – scale parameter, *b* – offset parameter, ψ – mother wavelet.

Continuous Wavelet Transform has many advantages but in speech signal processing its discrete form is more often used due to its simplicity [17]. Discrete Wavelet Transform (DWT) is defined by the following form [17]:

$$DWT(m,n) = \frac{1}{\sqrt{a^m}}\sum_n S(k)\psi(a^{-m}n - bk) \tag{3}$$

where: *S(k)* – indicates input signal, *a* (*a*>0) – scale parameter, *b* – offset parameter, ψ – mother wavelet.

One of the advantages of Discrete Wavelet Transformation, compared for example to Fourier transform, is that DWT provides accurate and uninterrupted time information, which is a significant enhancement for signal processing [18].

Two types of mother wavelet were considered in conducted researches: the Haar Wavelet and the Mexican Hat Wavelet. The first one is defined as follows [19]:

$$\psi(t) = \begin{cases} 0, for\ t < 0 \\ 1, for\ 0 \le t < 0{,}5 \\ -1, for\ 0{,}5 \le t < 1 \\ 0, for\ t \ge 1 \end{cases} \tag{4}$$

The Mexican Hat Wavelet has the following form [20]:

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{\frac{-1}{4}}(1-t^2)e^{\frac{-t^2}{2}}. \tag{5}$$

Sample graph of above-mentioned function is shown the in Figure 2.

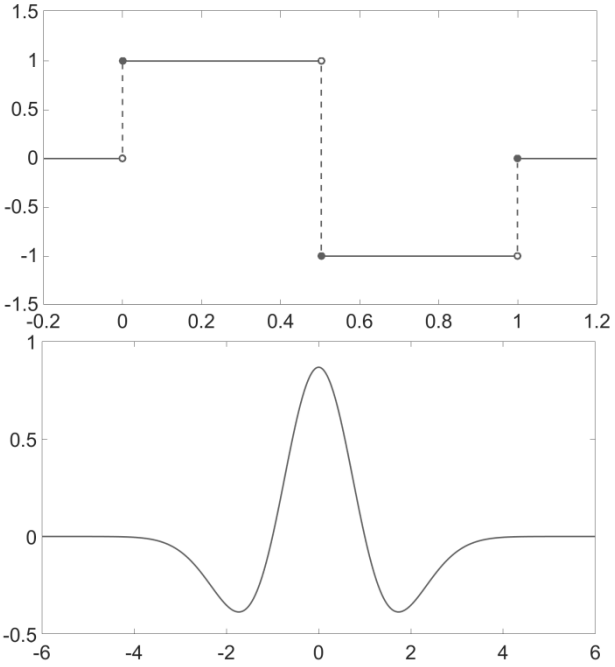

Fig. 2. Examples of wavelet functions: Haar (top) and Mexican Hat (bottom).

4.2 Speech signal processing scheme

The biggest challenge in conducted researches were preparation and processing of data process. It was divided into several steps. The first one related to noise reduction process from input signals. To fulfil this task The VOICEBOX: Speech Processing Toolbox for MATLAB was used. The second one related to speech signal values normalisation process to the range [-1,1]. These two steps constituted the pre-processing phase. In the next step the scaleogram was created. To create speech signal spectrum, the discrete wavelet transform was used. The example of signal decomposition is shown in the Figure 3.

In the next step, features were extracted from scaleograms. The main issue of this phase was the preparation of an input vector for processing by Kohonen Networks. At the beginning of features extraction process all scaleograms were transformed into grey scale images. The next step was a transformation to binary images which was as follows: all values, above specific threshold, were replaced by 1, other ones by 0. To define the threshold all values between 50 and 200 were tested. The best results were obtained at 100. The next step included scaleogram division into several subareas. The number of subfields depended on speech signal decomposition's level. The example of scaleogram division is shown in the Figure 4.

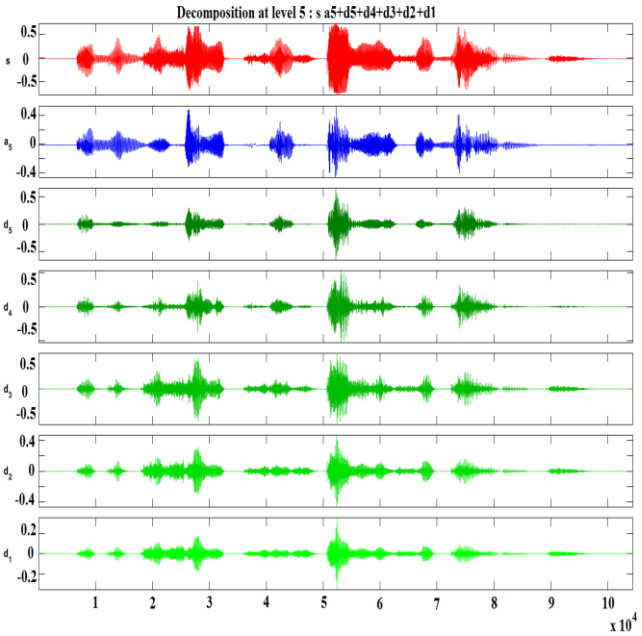

Fig. 3. Examples of speech signal decomposition using DWT.

Commonly used k-Nearest Neighbours algorithm was used also in this research, to determine the best level of signal decomposition. The fastest form of k-NN algorithm is its 1-NN version. In this algorithm, the unknown tested sample should have been assigned to the same group as its closest neighbour. Results of researches conducted during Statlog project [21], in which a several classifiers were compared, showed that for 75% of k-NN test the best results were achieved for 1-NN version. In that research the aforementioned algorithm was also used. Achieved results were shown in the Figure 5. Based on abovementioned algorithm, the 7^{th} and 9^{th} speech signal decomposition's level were used. It can be easily noticed that achieved results, under 7^{th} level of decomposition, were unsatisfactory. What is more, the processing time above 9^{th} level has been growing rapidly. The number of scaleogram subareas were multiple of level of decomposition. For 7^{th} decomposition's level the following numbers of subfields were tested: 70, 105, 140, 175, 210, 245, 270, 315, 350 and 385 and for 9^{th} decomposition's level: 90, 135, 180, 225, 270, 315, 360, 405.

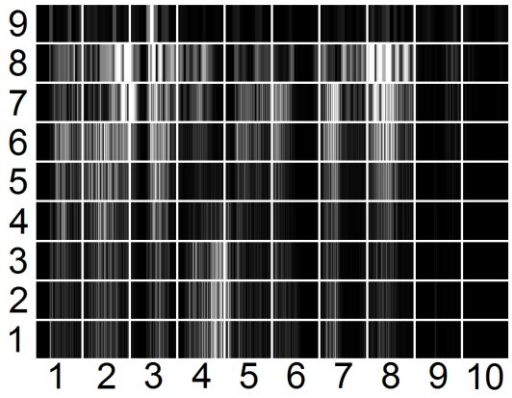

Fig. 4. Examples of scaleogram divisions.

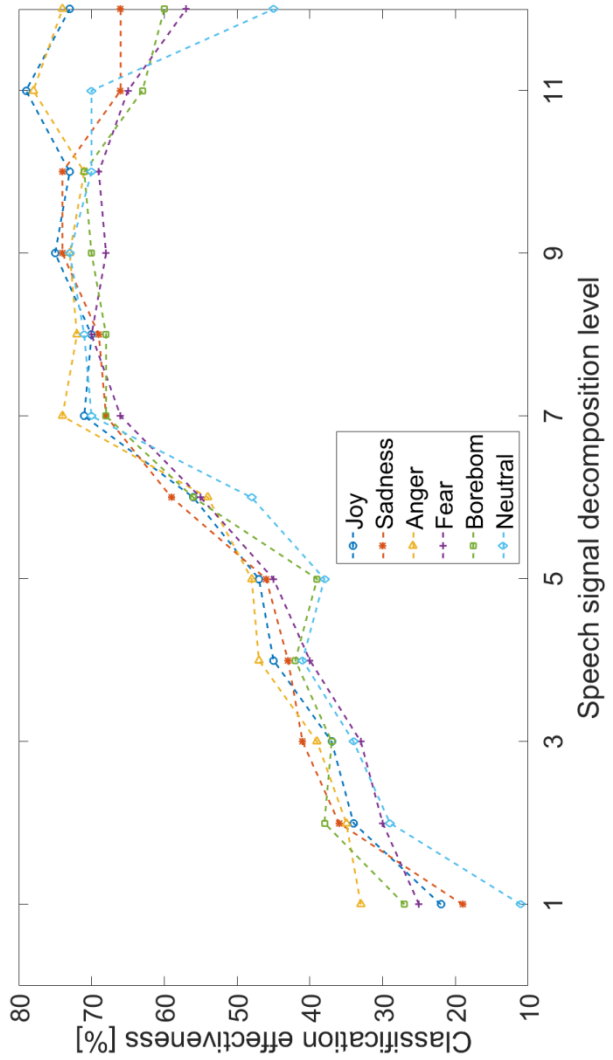

Fig. 5. Classification results using 1-NN rule depending on signal decomposition's level.

The best effectiveness was achieved from division into 210 subareas for 7^{th} decomposition's level and into 360 for 9^{th} level. The last step in Polish emotional speech signal processing, before classification, was summarising values in each subarea. The whole processing scheme is illustrated in the Figure 6.

4.3 Kohonen Networks Architecture and achieved results

Kohonen Networks belong to a group of Self-Organizing Maps. The main difference between this kind of neural networks and one-way artificial neural networks is that the correct output cannot be defined before training process (*a priori*). The main aim for Kohonen networks is to organise multidimensional information in such a way that it can be presented and analysed in space with smaller number of dimensions. As it was mentioned, scaleograms have been divided into 70, 105, 140, 175, 210, 245, 270, 315, 350 and 385 subfields for 7^{th} signal decomposition's level and 90, 135, 180, 225, 270, 315, 360, 405 for 9^{th} level of decomposition. In researches, a two-dimensional map was used, corresponding to scaleogram division size. The map in Kohonen network

was constructed in a way that for one input value there were four corresponding neurons in a map. So, the size of the map was four time bigger then input vector. The Kohonen networks work as follows:
1. Inputs relate to all nodes in map.
2. Each node stores a weight vector with identical size as an input vector.
3. Each node calculates its activation level as scalar product of vector weights and input vectors.
4. A node with the highest level of activation is the winner and can update its weight vector.
5. Nodes in the winner's neighbourhood may update their weight vectors.

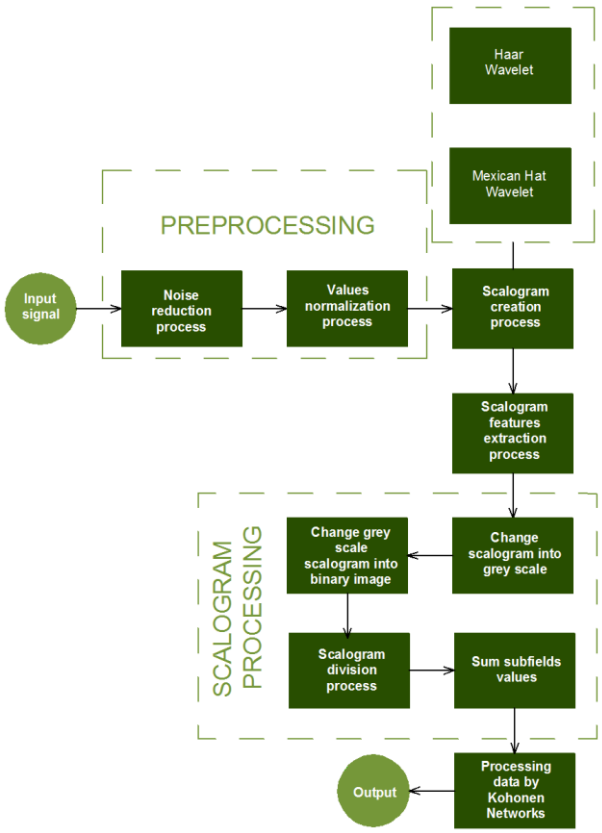

Fig. 6. Polish emotional speech signal processing scheme.

In conducted researches, a Euclidean distance weight function was used as a neighbourhood function. Number of epochs was set to 1000. As a training set 50% of recordings from the first database was used. Achieved results are presented in Tables 3 to 5.

Table 3. Achieved classification results for the first database.

Emotion	Correct classification [%]
Anger	72,5
Joy	77,4
Sadness	69,2
Neutral state	70,3
Boredom	72,7
Fear	79,8

Table 4. Achieved classification results for the second database.

Emotion	Correct classification [%]
Anger	74,7
Joy	76,5
Sadness	69,4
Neutral state	73,4
Boredom	77,1
Fear	77,6

Table 5. Achieved classification results for the third database.

Emotion	Correct classification [%]
Anger	77,4
Joy	76,6
Sadness	73,3
Neutral state	74,2
Boredom	73,5
Fear	80,2

It can be noticed that the strong emotions, such as fear or anger, were more often recognised correctly than the weak ones. What is more, regardless of used database, fear was identified with the highest efficiency and sadness with the worst. The justification for such results should be sought in the appearance of the speech signal spectrum which for sadness is quite similar as for boredom and neutral states; therefore, these three emotions were often confused. The opposite situation occurs in the case of fear, which spectrum was clearly different from spectrums of other emotional states. For all data the average effectiveness of classification was almost 76%.

5 Conclusions

Conducted researches showed that identification of emotion in speech signal is not trivial issue. There are not many publications directly related to the possibilities offered by spectrographic methods and Kohonen networks in issues connected to the identification of Polish emotional speech. Scaleogram division and determination of energy in its subfields allowed to create the appropriate input vector for Kohonen networks. Conducted researches and obtained results, as well as preliminary results of subsequent experiments, allow to assume that proposed method of speech signal processing can be universal and it is possible to uniquely identify the speaker's emotional state.

References

1. D. Ververidis, C. Kotropoulos, I. Pitas, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **149**, 1(2004)
2. P. Powroźnik, D. Czerwiński, Advances in Science and Technology Research Journal, **11**, 10, 32(2016)
3. S. P. Panda, A. K. Nayak, International Journal of Speech Technology, **2**, 9(2016)

4. D. Kamińska, A. Pelikant, IAPGOŚ, **10**, 03 (2012)
5. S. Ramakrishnan, Speech Enhancement, Modelling and Recognition – Algorithms and Applications, **7**, (2012)
6. D. Kamińska, T. Sapiński, D. Niewiadomy, A. Pelikant, Studia Informatica, **5**, 34 (2013)
7. T. Dalgleish, Nature Reviews Neuroscience, **18**, 5, 7(2004)
8. L. Zhen, G. Zhi, Affective Computing and Intelligent Interaction, **81**, (2005)
9. K. Scherer, Speech Communication, **13**, 40, 1-2(2003)
10. A. Janicki, M. Turkot, Przegląd telekomunikacyjny - wiadomości telekomunikacyjne, **98**, 8-9(2008)
11. J.H Conolly, E. A. Edmonds, J. J. Guzy, S. R. Johnson, A. Woodcock, International Journal of Man-Machine Studies, **7**, 24, 6(1986)
12. I. Józefczyk, Problemy Transportu, **7**, 4, 3(2009)
13. F. Burkhart, A. Paesche, M. Rolfes, F. Sendlmeier, B. Weiss, *A database of german emotional speech* (Proceedings of Interspeech 2005)
14. Database of Polish Emotional Speech, available: http://www.eletel.p.lodz.pl/bronakowski/med_catalog/ (Accessed 10.08.2014)
15. V. Maz, G. Schmidt, Applied and Computational Harmonic Analysis, 1, 6, 3(1999)
16. L. Zsu-Hsin Chuang, L. Wu,J. Wang, Sensors, **60**, 13, 8(2013)
17. A.Anoop Suraj, M. Francis T.S.Kavya T.M.Nirmal, Journal of Electrical Systems and Information Technology, **8**, 1, 1(2014)
18. R. Sharma, R.B. Pachori, U. R. Acharya, Entropy, **5**, 17, 8(2015)
19. U. Lepik, Applied Mathematics and Computation, **33**, 176, 1(2006)
20. Y. Hong, H. Liang, Optical Letters, **39**, 31, 3 (2006)
21. R. D. King, D. Feng, A. Southerland, Applied Artificial Inteligence, **2**, 9, 3(1995)