# Combination of Fourier and wavelet transformations for detection of speech emotions

Mariusz Ziółko, Paweł Jaciów, Magdalena Igras
Faculty of Computer Science, Electronics and Telecommunications
AGH University of Science and Technology
Kraków, Poland
{ziolko, migras}@agh.edu.pl, jaciow@student.agh.edu.pl

***Abstract*— The paper presents an approach to automatic recognition of emotions in speech signals. The applied method bases on the composition of two discrete frequency transformations. The wavelet transform was calculated first and next the Fourier transform was applied. The Fourier-wavelet transform representation is used to find the differences between emotions in speech signals. A set of approximately 30 seconds long speech signals was used to verify the efficiency of presented methods. It gives the possibility of analyzing the performance of speech emotion recognition in the Fourier-wavelet domain.**

***Keywords—speech emotions; wavelet transform; FFT***

## I. INTRODUCTION

The detection of non-linguistic content of speech signal, like speaker emotion, mood or attitude, gained an increasing attention in recent years. The voice timbre and a manner of speaking carry information which is complementary to the semantic content of speech but it is lost in standard speech to text transcription. The recognition of the non-verbal information does not only enrich the meaning of the words being spoken, but can also bring significant information on the speaker state and intention.

The vocal emotions recognition has become an important part of the affective computing technology [1], thanks to the growing importance of voice interfaces. For the needs of voice human-computer interfaces, automatic detection of user emotions can directly help in making the systems more adaptable, increasing user satisfaction and, at the same time, improving the application effectiveness. Likewise, Interactive Voice Response (IVR) systems can be supported with emotion detection. Also development of emotional robots and virtual reality systems will benefit greatly from judging users attitude and adapt according to the feedback. In the field of human-computer interaction cognitive aspects, investigation of vocal cues of emotions can also contribute to creating more naturally sounding emotional voices in speech synthesis systems.

From the speech technology point of view, the research on emotions in speech is relevant for speech and speaker recognition systems. Their efficiency is decreased by the variability of the way of speaking caused by emotions. Proper normalization techniques will compensate the differences and make the systems more robust.

The results of research can be also applied in other fields. In medical field, as a non-invasive method, it might be used in monitoring psychic disorders therapy (autism, bipolar disorder). In forensic psychology, detection of emotion from voice is crucial especially in the cases of absence (e.g. phone calls) of any other evidence. The authors of this paper intend to apply information about the speaker emotional state to create a caller profile in the emergency phone.

Vocal expression of emotions, even when only speech modality is available, is spontaneously recognizable by human perception with about 50-70% accuracy. Expression of basic emotions (happiness, sadness, anger, fear, surprise and disgust) is considered to be culture and language-independent phenomenon. Therefore, it is possible to extract acoustic features characteristic for affective states according to the paradigm of Scherer, assuming that each basic emotion can be described by a unique pattern or configuration of the acoustic parameters [2].

Automatic recognition of emotions in speech is a challenging task for several reasons. First of all, emotions in real life situations are usually coexisting and appear as complex configurations of basic emotional states. Secondly, emotions are multimodal phenomena, and some of them are more clearly expressed by facial mimics, gestures or physiological reactions. Hence, only some information on emotions is conveyed by voice. For example, positive emotions use to be expressed rather by visual cues, while the negative - by acoustic ones. Furthermore, the way of expressing emotions depend on speaker individual features, like their personality or sociologic characteristics. Also the listener ability to recognize emotions in speech is the effect of individual empathy and sensibility level. It can influence the results of human perception test during labeling of the recordings.

It is necessary to mention the importance of the proper database choice. The most common approach is using corpora of emotional speech simulated by actors. The contemporary tendency is to look for sources of spontaneous speech, although they cannot ensure such regularity as actors do. The emotions can be elicited with affective stimuli. The most natural sources of emotional speech, with respect to spontaneity and naturalness, are real-life situations, like call centers conversations, emergency phone calls, television live coverage. For this study, a corpus of acted emotional speech (described in chapter III) was used, while the further validation of the introduced method for the future applications will be based on real-life recordings from emergency phone.

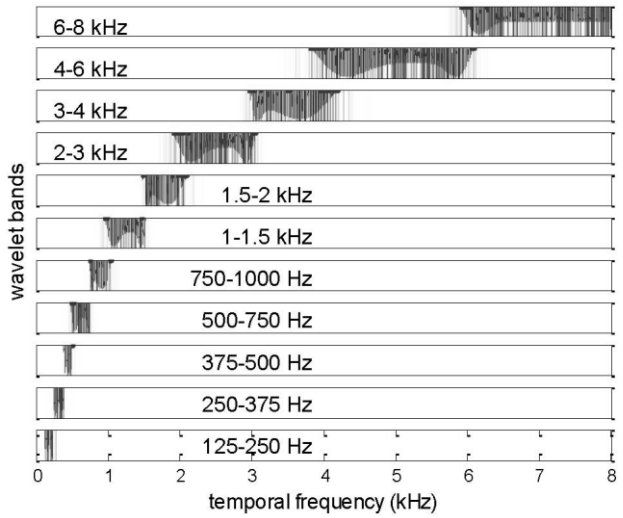



Fig. 1. Width of frequency subbands in speech spectra [11]

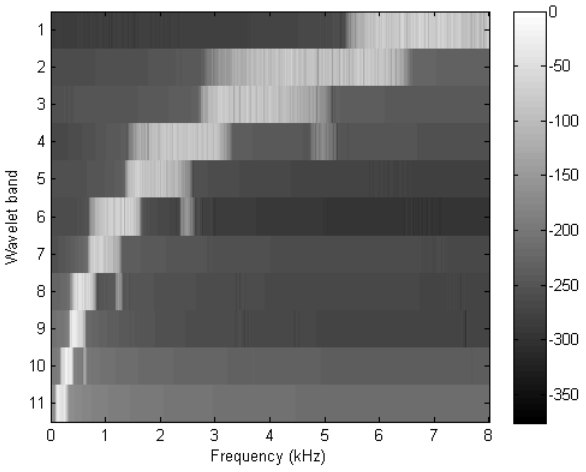



Fig. 2. An example of DFWT amplitude spectrum (speaker 10, fear)

Different vocal features correlated with emotions can be extracted from speech signal by a number of parameters in both, the time and frequency domain. Most often considered are: prosodic features, with focus on pitch (range, standard deviation, slope), energy and vocal stress. Also speech rate and pauses (including duration and number of silent, breath or filled pauses) are significant indicators of emotions.

The original representation in the time domain usually gives little information about the speech signal properties. To make information more noticeable it is necessary to use some transformation. The efficiency of speech technology algorithms usually strongly depends on the choice of transformation. The goal of our work is to analyse, if the Discrete Fourier-Wavelet Transform (DFWT) meets the demand of speech emotion detection.

Spectral features of speech signal (e.g. mel-cepstral coefficients [3] or wavelet transforms [3,4,5]) were considered alone or in fusion with prosodic features [4,5]. In our previous work [6], we described a method based on energy values in frequency subbands obtained from the Discrete Wavelet Transform (DWT) with perceptual (mel) scale. DWT was tested also in [11]. In another approach [8], experiments on applicability of DWT, Wavelet Packets Transform (WPT) and Perceptual Wavelet Packet Transform (PWPT) are performed, where the best results were achieved for the DWT method. Emotion modelling and emotion conversion using transition in subbands of WPT is performed in [9]. The fusion of Fourier and DWT has been investigated in the realm of emotion detection and some results are presented in this paper.

As classifiers, the most frequently are used: K - Nearest Neighbours (K-NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), Back-Propagation Neural Networks (BPN), Binary Decision Trees, Gaussian Mixture Models (GMM), Linear Discriminant Analysis (LDA) [4,5]. The wavelet energy coefficients were used for feature extraction for neural networks [11] or SVM classifiers [8]. All these methods achieved recognition efficiency about 60-80% [4, 5].

The detection system should be independent of spoken text and should be based on the speech characteristics. The aim of speech emotion detection is to provide an efficient and accurate algorithm to distinguish the individual properties of each emotion contained in speech signal.

The paper is organized as follows: section II introduces the mathematical background of the DFWT, next the database is described in section III. Section IV contains specification of our experiments and presentation of results. The paper is concluded in section V.

## II. DISCRET FOURIER-WAVELET TRANSFORMATION

Discrete Wavelet Transform (DWT) is the new generation of mathematic tool [10] which belongs to the group of frequency transformations and is used to obtain a time-frequency spectra of signals. To start DWT calculations, values of a discrete signal $s(n)$ are identified with coefficients $c_{m+1,n}$, by substitutions

$$c_{m+1,n} = s(n), \tag{1}$$

where $m+1$ is large enough resolution level to obtain the assumed frequency band. The coefficients of the lower level are calculated by applying the well-known [10] formulae

$$c_{m-1,n} = \sum_k h_{k-2n} c_{m,k} \tag{2}$$

$$d_{m-1,n} = \sum_k g_{k-2n} c_{m,k}, \tag{3}$$

where $h_m$ and $g_m$ are the constant coefficients which depend on the arbitrary chosen scale function. The coefficients of next resolution levels are calculated recursively by applying formulae (2) and (3). In this way values

$$DWT = \left\{ \left\{ d_{m,n} \right\}_n, \cdots, \left\{ d_{0,n} \right\}_n, \left\{ c_{0,n} \right\}_n \right\} \quad (4)$$

for *m*+2 levels are obtained. Values $d_{m,n}, ..., d_{0,n}$ and $c_{0,n}$ have two arguments and represent a wavelet spectrum of signal $s(n)$ while parameter *m* called resolution level correlates with the frequency bands. Parameter *n* is a discrete time, hence wavelet spectrum (4) carries both, time and frequency analysis. The number of samples $N_m$ depends on resolution level *m*. Generally each higher resolution level has about twice more samples than the next lower resolution level. Under assumption that $N_m$ is an even number, the condition $N_{m-1} = N_m / 2$ is fulfilled.

The classic discrete decomposition schemes are dyadic and do not provide sufficient number of frequency bands for effective speech analysis. In case of the perceptual scale the number of subbands must be increased. Wavelet packets provide more frequency bands and the decomposition structure which provides a perceptual frequency analysis is suggested in [11,12]. It gives more frequency bands than dyadic wavelet decomposition and less than uniform frequency distribution. Such solution seems to be a good compromise and gives frequency decomposition inspired by the widely use mel scale. This goal was reached by the perceptual frequency division for the desired frequency bands.

It enables adaptation of the time-frequency analysis to particular speech signals properties. Vectors $\left\{ d_{m,n} \right\}_n$, which constitute a part of spectrum (4), should be split into two vectors $\left\{ e_{m,n} \right\}_n$ and $\left\{ f_{m,n} \right\}_n$ to represent the additional frequency bands. Spectra for the eleven required frequency subbands are computed by applying procedures

$$f_{m,n} = \sum_k h_{k-2n} d_{m,k} \quad (5)$$

$$e_{m,n} = \sum_k g_{k-2n} d_{m,k} \quad , \quad (6)$$

where $2 \le m \le 6$. Such case is presented in Fig. 1 and was used in the analysis described below. From the acoustic point of view, eleven subbands seem to be the best frequency representation of the speech properties in terms of speech analysis [11]. The information about the lowest frequency band, from 0 to 125 Hz, is represented by a vector $\left\{ c_{0,n} \right\}$. This part of DWT was skipped in the spectral representation because it carries a relatively strong noise and little information about speech properties. Finally, spectra

$$WP = \left\{ \left\{ f_{6,n} \right\}_n, \left\{ e_{6,n} \right\}_n, ....., \left\{ f_{2,n} \right\}_n, \left\{ e_{2,n} \right\}_n, \left\{ d_{1,n} \right\}_n \right\} \quad (7)$$

were computed to analyze speech signals.

The Fourier-wavelet transform combines two well known transforms, by performing Fourier transform along the translation parameter of wavelet packet (7). For the case of digital speech signal (1), we compute

$$DFWT = \left\{ \left\{ \hat{f}_{6,k} \right\}_k, \left\{ \hat{e}_{6,k} \right\}_k, ....., \left\{ \hat{f}_{2,k} \right\}_k, \left\{ \hat{e}_{2,k} \right\}_k, \left\{ \hat{d}_{1,k} \right\}_k \right\} \quad (8)$$

by applying separately the Fast Fourier Transform (FFT) to each frequency band of wavelet packet transform (7) and $k = 1,..., N_b$. Values of *DFWT* are complex and their modules $|DFWT|$ are parameters which give the specific and individual frequency characteristics for voice of each speaker.

The sampling frequency has been set to 16 [kHz]. The *DFWT* were calculated for all eleven resolution levels. An example of spectrum, with logarithmic scale for its module values, is presented in Fig. 2 and the high-resolution levels are in the upper part of this plot.

## III. DATABASE

A part of Polish emotional speech database [13] was used for the experiments. The corpus contains good quality (.wav files of 16 bit, 16 kHz, SNR > 30 dB) recordings of actors and theatrical art students, reading the same content each time with one of the following emotions: anger, fear, sadness, joy, surprise and neutral state. For our analysis we selected the recordings of approximately 30 seconds of continuous speech of 6 male (including 4 actors) and 5 female speakers (including 3 actors). The semantic content was the same for each emotion: an informative text of neutral meaning.

## IV. EXPERIMENTS

The workflow of algorithm is presented in Fig. 3. At the beginning the speech signals of all speakers were normalized in reference to the square root of energy, according to the formula

$$\bar{s}(n) = \frac{s(n)}{\sqrt{\sum s^2(n)}} . \quad (9)$$

Then, for each normalized recording, *DFWT* were computed, as it was described in Section II.

Number of samples of each $|DFWT|$ subband was reduced by counting sums of adjacent 500 samples. Thanks to this, the average $|DFWT|$ values $\left\{ \overline{w}_b(i) \right\}_{i=1}^{I}$ were obtained. In each frequency subband the number of elements was reduce to $I = \lfloor N_b / 500 \rfloor$ (the examples are presented in Fig. 4).

The role of the next step is to fit a curve for $\left\{ \overline{w}_b(i) \right\}_{i=1}^{I}$ values clearly greater than zero. The role of arbitrary chosen curve is to substitute many points by at the most four parameters. Asymmetrical distribution of points excluded the Gauss approximation. The curve fitting was performed by Padé approximation

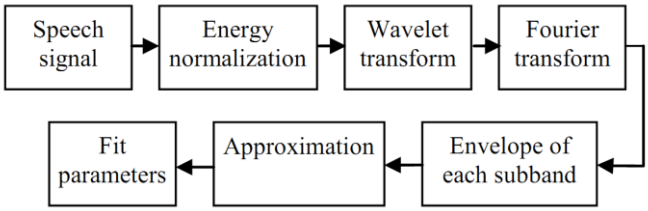
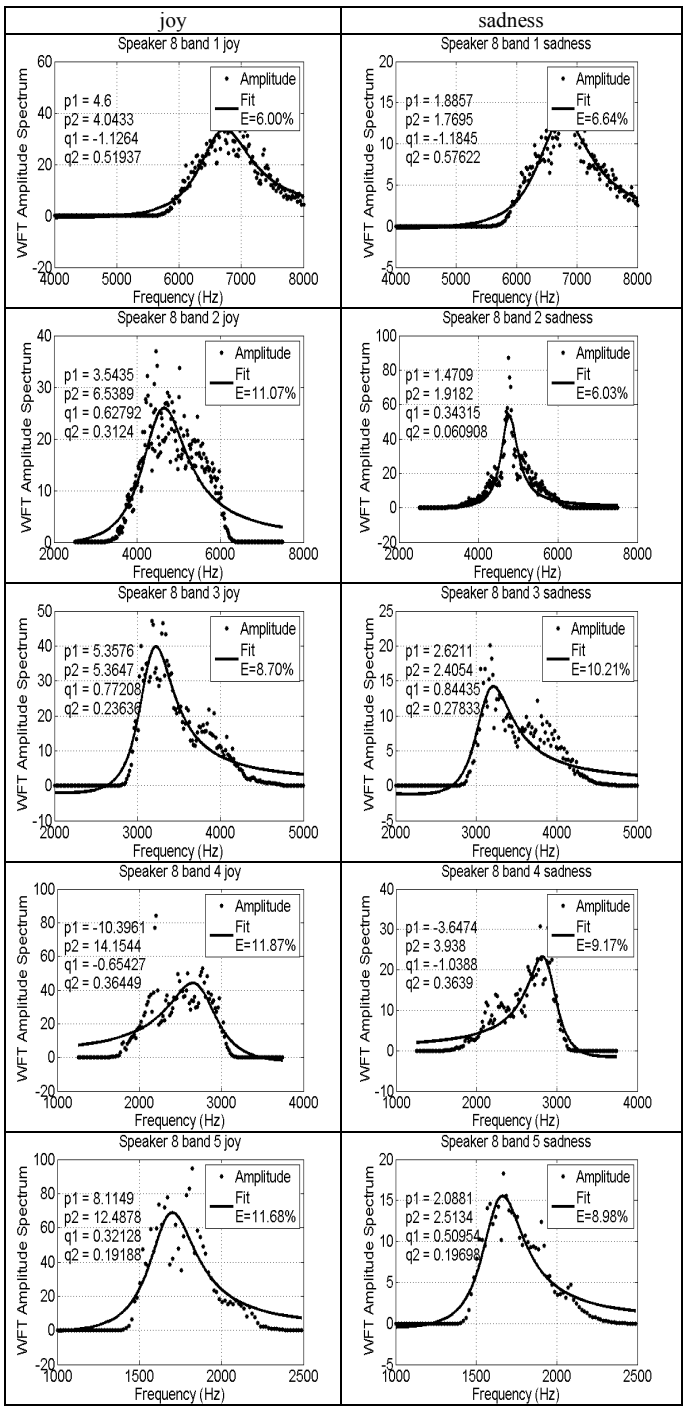



Fig. 3. Workflow of the parametrization algorithm



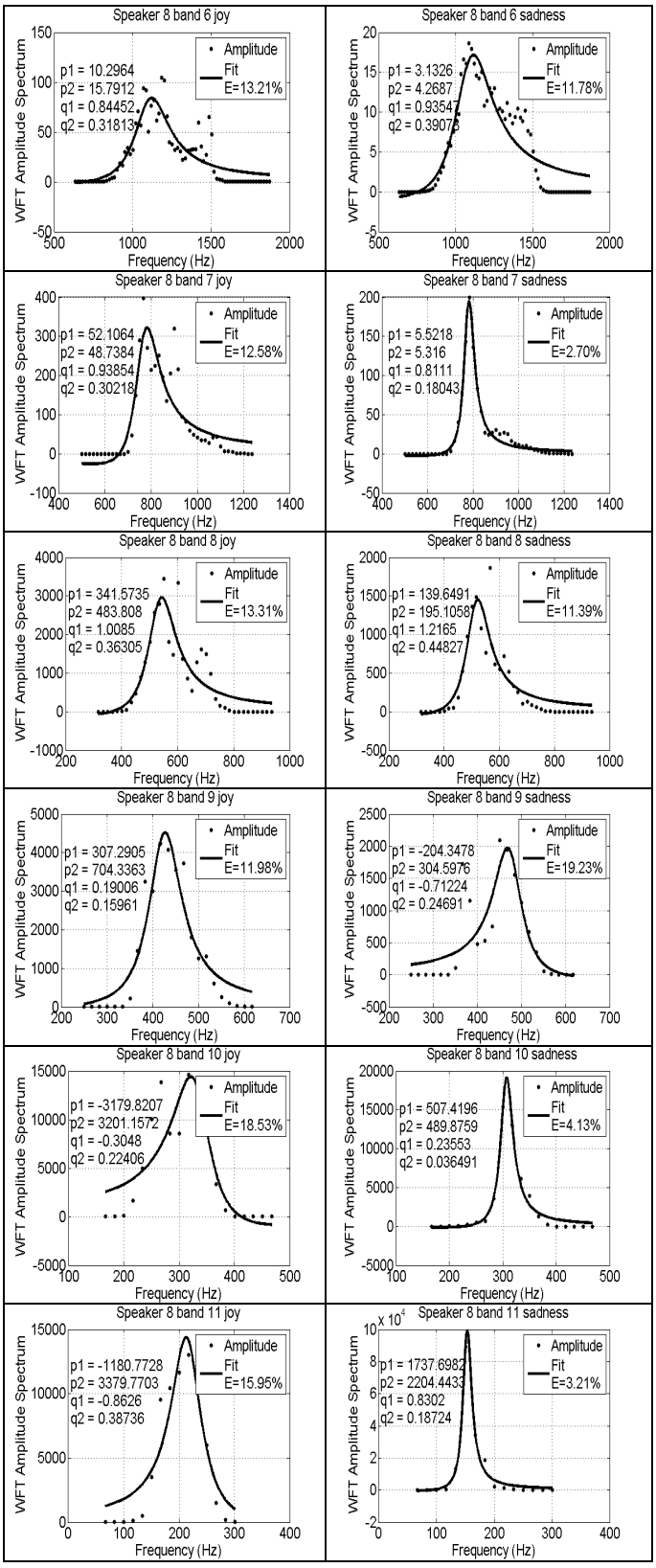




Fig. 4. Comparison of two emotions for the same subbands and the same speaker (female)

$$w_b(i) = \frac{p_1 i + p_2}{i^2 + q_1 i + q_2}, \tag{10}$$

where $p_1, p_2, q_1, q_2$ are the parameters describing each subband. Padé approximation (10) was fitted for each frequency band of each emotion and for each speaker. The values of goodness of fit for the Padé approximations (presented in Fig.4) are the normalized root-mean-square errors

$$E_b = \frac{\sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(\overline{w}_b(i) - w_b(i)\right)^2}}{\max_i \overline{w}_b(i) - \min_i \overline{w}_b(i)} 100\% \cdot \tag{11}$$

As a result, each recording is characterized by the vector containing 44 parameters = 4 Padé coefficients × 11 subbands. Finally, for every emotion the mean vectors (each consisting of the 44 parameters) were calculated, separately for female and male speakers. Let us name this set of emotion parameters as the characteristic vector. It constitutes a model for each emotion contained in speech signal and can be used to detect which kind of emotion appears in verified speech.

To detect the emotion, the speech parameters should be compared with parameters of models obtained for emotions: joy, sadness, fear, surprise, anger and eventually neutral state. Some mathematical metrics (e.g. Euclidean) can be used to find the nearest characteristic vector. From among two tested criteria (Euclidean distance and the inner product), the more useful in recordings classifying were the inner products of its 44 parameters and the characteristic vectors for all emotions. The emotion for which the inner product obtains the greatest value is recognized as the best suited emotion.

An example results, in the form of the confusion matrix are presented in Tab. I. For these two examples, joy and sadness for Female 5 and joy and surprise for Male 2, were detected correctly, as the inner products obtained the greatest values for the proper emotions. For other cases, the proper emotions were located on the second or sometimes on the third positions. The results are gathered in Tab. II. It is worth to notice that "sadness" was always perfectly detected for all women voices. The worst results, from presented in Tab. II, were obtained for the male emotion "joy", where the average position is a bit lower than the second one.

## V. CONCLUSIONS

The preliminary study of a novel fusion of Fourier and wavelet transform applicability in the field of emotion detection was presented. In the proposed model, features are extracted from the Padé approximation of the amplitude shape in the $|DFWT|$ frequency subbands.

The results, obtained from the experiments of actors' emotional speech, showed that it is possible to observe and to make use of some characteristic irregularities which are not directly detectable in either the wavelet or Fourier spectrum. The Fourier-wavelet analysis seems to be a promising tool to distinguish the emotions included in speech signals. We noticed that the most significant differences between emotions appeared in medium subbands (see Fig. 4), especially 3 and 4 (500 - 750 Hz). The best results for women recordings were achieved for detection of joy and sadness. For men our method was the most distinctive in case of joy and surprise.

Although using acted emotional speech for measuring acoustic correlates of emotions, has been widely criticized for overemphasized and stereotypical expressions, it was a useful material to verify performance of DFWT tool in the task of discrimination of different manners of speaking.

Further research in this field will include modelling the real-life emotions, using emergency phone calls database. We will explore also other aspects of the Fourier-wavelet transform potential for measuring vocal correlates of emotions (e.g. computation of more features to capture additional cues of emotions revealed by the DFWT).

TABLE I. INNER PRODUCTS OF CHARACTERISTIC VECTORS WITH AVERAGE CHARACTERISTIC VECTORS APPROPRIATE FOR EACH SPEAKER SEX

| | Female 5 | | Male 2 | |
|---|---|---|---|---|
| | Joy | Sadness | Joy | Surprise |
| Neutral | 83 | 73 | 85 | 98 |
| Joy | **136** | 30 | **111** | 147 |
| Sadness | 71 | **120** | 62 | 137 |
| Fear | 110 | 32 | 55 | 147 |
| Surprise | 90 | 55 | 104 | **203** |
| Anger | 88 | 29 | 96 | 138 |

TABLE II. POSITION OF PROPER DETECTION

| | joy | sadness |
|---|---|---|
| Female 1 | 2 | 1 |
| Female 2 | 1 | 1 |
| Female 3 | 1 | 1 |
| Female 4 | 1 | 1 |
| Female 5 | 1 | 1 |
| **Average** | **1.2** | **1** |
| | joy | surprise |
| Male 1 | 3 | 1 |
| Male 2 | 1 | 1 |
| Male 3 | 3 | 2 |
| Male 4 | 2 | 1 |
| Male 5 | 2 | 2 |
| Male 6 | 3 | 1 |
| **Average** | **2.3** | **1.3** |

## ACKNOWLEDGEMENT

The project was supported by the National Research and Development Centre granted by decision 072/R/ID1/2013/03.

## REFERENCES

[1] R. Picard, "Affective Computing," MIT Technical Report no 321, 1995.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," Speech Communication, vol. 40, pp. 227– 256, 2003.

[3] K.V. Kishore and K.P. Satish, “Emotion recognition in speech using MFCC and wavelet features,” IEEE 3rd International Advance Computing Conference (IACC), 22-23 Feb. 2013, pp.842 -847.

[4] S. G Koolagudi and K.S. Rao, “Emotion recognition from speech: a review,” International Journal of Speech Technology, Vol. 5, No 2, pp. 99-117, 2012.

[5] T. Vogt, E. André, and J. Wagner, “Automatic recognition of emotions from speech: A Review of the literature and recommendations for practical realisation,” Lecture Notes in Computer Science, Vol. 4868, 2008, Affect and Emotion in Human-Computer Interaction, pp. 75-91.

[6] M. Igras, M. Ziółko, J. Gałka, “Wavelet evaluation of speaker emotion,” Proceedings of the Eighteenth National Conference on Applications of Mathematics in Biology and Medicine, Krynica Morska, 23–27 September 2012, pp. 54–59.

[7] A. Firoz, A. Raji Sukumar and P. Babu, “Discrete wavelet transforms and artificial neural networks for speech emotion recognition,” International Journal of Computer Theory and Engineering, vol. 2, no. 3, pp.1793-8201, June 2010.

[8] S. Emerich and E. Lupu, “Improving speech emotion recognition using frequency and time domain acoustic features,” Proceeding of SPAMEC 2011, Ckuj-Napoca, Romania.

[9] V. A. Degaonkar and S.A. Apte, “Emotion modeling from speech signal based on wavelet packet transform,” International Journal of Speech Technology, Springer, vol. 16, issue 1, pp 1-5, 2013.

[10] I. Daubechies, Ten lectures on wavelets. SIAM, 1992.

[11] M. Ziółko, J. Gałka, B. Ziółko and T. Drwiega, “Perceptual wavelet decomposition for speech segmentation,” Proceedings of the Interspeech, Makuhari 2010, pp. 2234-2237.

[12] O. Farooq and S. Datta, “Mel Filter-like admissible wavelet packet structure for speech recognition,” IEEE Signal Processing,Letters, vol. 8, pp. 196-198, 2001.

[13] M. Igras, B. Ziółko, Baza danych nagrań mowy emocjonalnej (Eng. Database of emotional speech recordings), Studia Informatica, vol. 34 no. 2B, pp. 67–77, 2013.