jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2012, VOL. 58, NO. 2, PP. 165–170
Manuscript received August 24, 2011; revised May, 2012. DOI: 10.2478/v10177-012-0024-4

Recognition of Human Emotion from a Speech Signal Based on Plutchik's Model

Dorota Kamińska and Adam Pelikant

***Abstract*—Machine recognition of human emotional states is an essential part in improving man-machine interaction. During expressive speech the voice conveys semantic message as well as the information about emotional state of the speaker. The pitch contour is one of the most significant properties of speech, which is affected by the emotional state. Therefore pitch features have been commonly used in systems for automatic emotion detection. In this work different intensities of emotions and their influence on pitch features have been studied. This understanding is important to develop such a system. Intensities of emotions are presented on Plutchik's cone-shaped 3D model. The k Nearest Neighbor algorithm has been used for classification. The classification has been divided into two parts. First, the primary emotion has been detected, then its intensity has been specified. The results show that the recognition accuracy of the system is over 50% for primary emotions, and over 70% for its intensities.**

***Keywords*—emotion detection, Plutchik's wheel of emotion, speech signal,**

I. Introduction

In connection with continuous development of technology, scientists try to figure out new solutions supporting the relation between human and machine. Researchers devise methods of speech recognition. However, communication between human and machine is still affected and not efficient enough. Conversation between humans contains information about the emotional state of a speaker. Therefore, detecting emotions in human–machine communication is gaining more and more attention in the speech research community. Emotion recognition system could widely improve cooperation with machines and allow them to react appropriately to human intentions. The goal here is to create this kind of system.

This paper refers to the psycho-evolutionary theory of Robert Plutchik, which has ten main postulates [1]:

- The concept of emotion is applicable to all evolutionary levels, applies to humans and animals.
- Emotions have an evolutionary history and have developed various forms of expression in different species.
- Emotions have an adaptive role in helping organisms to survive the threat posed by the environment.
- Despite the differences in the forms of emotional expression in different species, there are certain common elements, or general patterns, that can be identified.
- There is a small number of basic emotions.

Dorota Kamińska is a scholarship holder of project entitled "Innovative education ..." supported by European Social Fund.

D. Kamińska and A. Pelikant are with the Institute of Mechatronics and Information Systems, Technical University of Łódź, Stefanowskiego 18/22, 90-924 Łódź, Poland (e-mail: dorota.kaminska@p.lodz.pl).

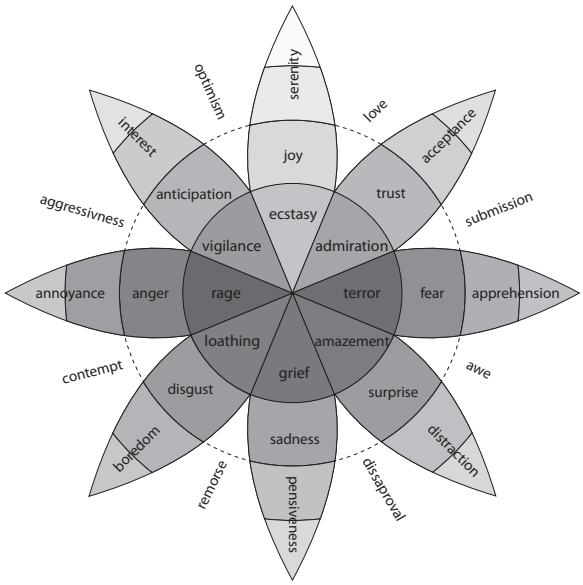

Fig. 1. Plutchik's model of emotion.

- All other emotions (mixed or derivative states) occur as combinations, mixtures, or compounds of the primary emotions.
- Primary emotions are hypothetical constructs, a kind of idealized states whose properties and characteristics can only be inferred from various kinds of evidence.
- Primary emotions can be conceptualized in terms of pairs of polar opposites.
- All emotions vary in their degree of similarity to one another.
- Each emotion can exist in varying degrees of intensity at different levels of arousal.

Plutchik created a model of emotions, which describes postulates of his theory, presented in Fig. 1. He suggests that there are eight primary emotions related biologically to the adaptation for survival: joy versus sadness, anger versus fear, trust versus disgust and surprise versus anticipation. From the merger of primary emotions more complex emotions are formed. Furthermore, primary emotions can be expressed at different intensities [2]. For example the group related to anger is represented by rage, anger and annoyance. This paper presents how intensity of different emotional states affects fundamental frequency of human voice. Vocal emotions are recognized using k-NN classifier based on statistical features extracted from an utterance

The remainder of the paper will be organized as follows. Section II will present an overview of the literature related

to the recognition of emotions. Section III will describe the features utilized in this study. Section IV will present the used method of classification. Section V will detail the results of the emotional classification task. Finally, Section VI will provide concluding remarks and future work.

II. Related Work

Recently, literature related to emotion recognition has grown widely. There are a few ways to recognize emotion by machines. Existing systems are mainly based on image analysis. Sophisticated new camera systems can detect lies and emotional states from subtle changes of expression and the flow of blood to one's skin [3]. However, this kind of projects are still tested on willing volunteers rather than in real-life or high stakes situation. Other such systems are tested for audio-visual monitoring primarily to detect negative emotion for violence prevention in public places [4].

In terms of range of emotions, most researchers concern the detection of negative and positive emotions, without indicating a specific state [5] or focus on aggression detection [6]. There are also many studies investigating the so-called basic emotions like happiness, anger, suprise, sadness, fear, disgust and neutral state, which are most commonly used because of their availability in public databases. For the purpose of this project, wide range of basic emotions (primary emotions and their intensities) basing on Plutchick's model have been tested.

There are two most popular approaches in emotion recognition from speech: single and multi-dimensional approach. In the first one system detects one state from closed set of enumerated emotions [7]. In the second one system detects the intensity of some components (called primitives) of recognized emotions. The most common is three-dimensional analysis, which locates emotions in the space of valence, activation and dominance [8]. Sometimes the space is reduced to two dimensions (activation and evaluation) like in [9].

Analysis of speech signal is divided into three main parts: first speech emotional database creation, then features extraction and classification.

First important step is the database selection. Most of the research is based on recordings acted by professionals. This kind of recordings are usually precise and good quality. An example of such a database is Berlin emotional database, which is a standard for emotion detection. It contains 10 emotional utterances (5 short and 5 longer sentences) simulated by ten actors (5 females and 5 males), all spoken in German [10]. Some scientists collect data sets using movies, television sitcoms, call centers or the Internet [11] or even from infant cry [12]. Others create spontaneous speech databases. This kind of recordings are difficult to acquire, but contain very valuable material-collection of natural speech signal reflecting the emotional responses of the speakers [13]. Usually they are created by provoking an appropriate response or recording speakers in natural situations, such as talking with the medical dispatcher as in [14] or with psychologist [15]. Some affirm that sole use of audio database is insufficient. For example Wang and Guan carried out an experiment with audio and visual data. They have used many different features and different classification algorithms. They proved that combination of audio and visual gives better results then either alone [16], achieving results of 89.2% performance accuracy.

The main question in emotions detection, based on speech signal, is the choice of the appropriate features. The most common is heuristic approach – from the speech signal numerous parameters are extracted. Then this group of parameters is subjected to selection algorithm to choose the most discriminative descriptors. Generally scientists use prosodic features (pitch, timing, loudness and energy), which contain useful information for detecting systems [17]. However, systems based solely on prosodic features do not provide satisfactory results. Therefore more sophisticated features like values of formant frequencies [18], Mel-Frequency Cepstral Coefficients (MFCC) [19], Linear Predictive Coefficients (LPC) [20], Log Frequency Power Coefficients (LFPC) are used [21].

Classifier selection is as important as appropriate features extraction. From a list of simple statistical classifiers, usually k Nearest Neighbor with different metrics is used. Other scientists use more advanced classifiers like Support Vector Machines (SVM) [18], Dynamic Time Warping (DTW), Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM), which is most popular in emotion recognition. Finally some of the researchers use neural networks like back propagation neural network [22] or Self Organizing Maps (SOM). In some publications scientists compare effects of several different classifiers. Gaurav [23] has done a comparison between k-NN and Support Vector Machines.

One of the most advanced programs in this field of interest is the work of an Israeli company called eXaudios. Their computer program called Magnify decodes human voice to identify a person's emotional state. Some companies in the United States already use the system in their call centers. eXaudios is testing use of such systems in diagnosing medical conditions like autism, schizophrenia, heart disease and even prostate cancer [24]. They claimed that the tone of voice is universal and provides a solid basis for creating even a multi lingual emotion detection system, which is a serious challenge even in human-human conversation [25].

III. Features

Representation of the signal in time or frequency domain is a complex image. Therefore, the features are sought to determine signal properties. In this part extracted features will be presented.

A. Fundamental Frequency

During speech, vocal folds can be in two states: the rhythmic opening for a voiced sound or entirely open for an unvoiced sound. If the sound is a voiced response of the vocal tract is a periodic signal, which consists of Dirac delta series. Distance between impulses is the duration and its reverse is the fundamental frequency F0 [26].

The psychological correlate of F0 is pitch. It is an individual attribute, depends on the size of the larynx, tension and size of the vocal folds, age and gender of the speaker. For example a bass voice has a lower fundamental frequency than

a soprano. A typical adult male will have a fundamental frequency of from 85 to 155 Hz, a typical adult female from 165 to 255 Hz. Children and babies have even higher fundamental frequencies. Infants show a range of 250 to 650 Hz, and in some cases go over 1000 Hz. A 10 year old boy or girl might have a fundamental frequency around 400 Hz [27].

Fundamental frequency variation within a range of frequencies is a natural process in human speech. This is heard as the intonation pattern or melody. During singing the fundamental frequency of the singers voice is controlled according to the melody of a song. Nevertheless, a person's relaxed voice usually can be characterized by a natural fundamental frequency.

During speech the range of F0 changes in relation to intonation, which plays a major part in expressing emotion [28].

There are many methods to determine the fundamental frequency. In this paper F0 has been extracted using the autocorrelation method. The analysis window was set to 20 ms with 50% overlap. It is difficult to objectively assess the behavior of F0 based on the chart. Therefore, statistical parameters related to F0 have been extracted and presented in Table II.

IV. Classification

Classification is an algorithm, which assigns objects to groups, called classes, based on object features. The features values, which are a source of information about the object, are usually presented by a vector:

$$x_j = [x^1, x^2, ..., x^d], \tag{1}$$

where d is the number of features, x^k is a feature value. All feature values in a task are called the training set CU. The goal of classification is to assign a particular class for an individual object x_j.

A. KNN Algorithm

In k-NN algorithm the recognition process involves calculating distances in parameters space X between the unknown x_j object and all objects of the training set:

$$x_k \in CU, for\ k = 1, 2, ..., I, \tag{2}$$

where I is the number of training examples.

Various metrics are used to calculate the distance. In this studies the Manhattan distance, which is presented in (3), has been selected.

$$d(x_j, x_k) = \sum_{i=1}^{n} |x_j^i - x_k^i| \tag{3}$$

Obtained distances are sorted in an ascending order. Object x_j is assigned to this class, which is the most common among k nearest objects [29].

V. Experiments

Studies have been carried out according to the algorithm shown in Fig. 2. Main steps are described in the following subsections.

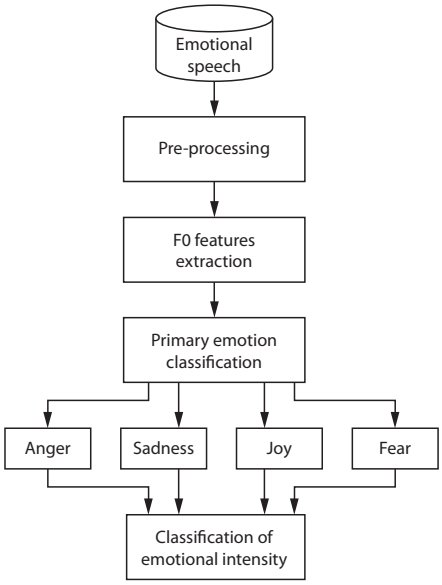

Fig. 2. Algorithm for audio signal processing.

TABLE I
Number of Used Utterances

Group of emotion	Intensities	Number
Anger	Rage	96
	Anger	98
	Annoyance	91
	All	285
Sadness	Grief	98
	Sadness	84
	Pensiveness	96
	All	278
Joy	Ecstasy	100
	Joy	95
	Serenity	80
	All	275
Fear	Terror	96
	Fear	98
	Apprehension	91
	All	285

A. Database Details

The main problem in emotion detection systems is the creation of an efficient database. For the purpose of these studies Polish emotional database has been created. It is divided into four groups represented by following primary emotions: anger, joy, fear and sadness. Each group consists of three different intensities. Thus, database consists of 12 emotional states. The recordings were taken from eight (four speaker per gender) healthy adult native Polish speakers of all ages. The first set consisted of 1950 utterances. The recordings were taken in an anechoic chamber, saved in PCM WAVE file format with 44100 Hz sampling rate. Based on the completed records, eight people made their classification into 12 groups (classes) of emotions. Selection of ambiguously-defined recordings was made this way. After selection process database has been reduced to 1123 unequivocal utterances as presented in Table I.

Proportional distribution of samples in each group has been preserved.

B. Pre-processing

The goal of pre-processing is signal adaptation for further processing and basic analysis. All collected utterances may contain background and microphone noise. Wavelet thresholding was used to the de-noising recorded utterances. Moreover, for further analysis, all collected utterances have been segmented into 20 ms frames using Hamming window with 50% overlap.

C. Features Extraction

Selection of efficient acoustic features is a critical point. It is quite difficult to create a not numerous vector, which describes the object of analysis well [30]. In this paper the influence of demonstrated emotional states on F0 contour has been presented. Following figures present typical F0 contours for four basic emotions and their intensities.

There are three anger intensities: rage, anger and annoyance For rage F0 increases noticeably in relation to neutral speech and also to its intensities. As Fig. 3 shows, this emotion appears to progress on a higher level in voice pitch. The lowest values were obtained for annoyance. Along with increase of emotion intensity the pitch range becomes much wider and its rises have a greater steepness. Particularly, pitch abruptness is significant on accented syllables.

According Plutchik's model joy has three intensities: ecstasy, joy and serenity. These vocal emotional states (similar with rage, anger and annoyance) characterized by increases in F0 mean, range and variability. However, pitch changes are smoother compared to the previous group. Although, increases are still proportional to the intensity of articulated emotion.

Grief, sadness and pensiveness have very similar F0 contours, also similar with the neutral speech. There is general decrease in F0 mean, range and variability and also downward-directed intonation contour. All of them are spoken with a small amount of change, F0 is almost constant. As in previous cases increases are proportional to the intensity of emotion.

The last group of emotion consists of terror, fear and apprehension. During the examination higher F0 mean and wider F0 range were found in comparison with neutral speech contour. Charts of all intensities are characterized by an initial increase and stabilize at the end of the utterance. The articulation tends to be precise. The effect of the intensity for the fundamental frequency is the same as in other emotional groups.

For the purpose of classification features described in Table II have been extracted from each signal (from the test and training sets).

Final vectors contain 24 features. Each vector has been subjected to standardization. The result of the standardization is a feature vector, which mean value is 0 and standard deviation value is 1. All the features have equal contribution to the value of the Manhattan distance, which is computed in the next step – classification.

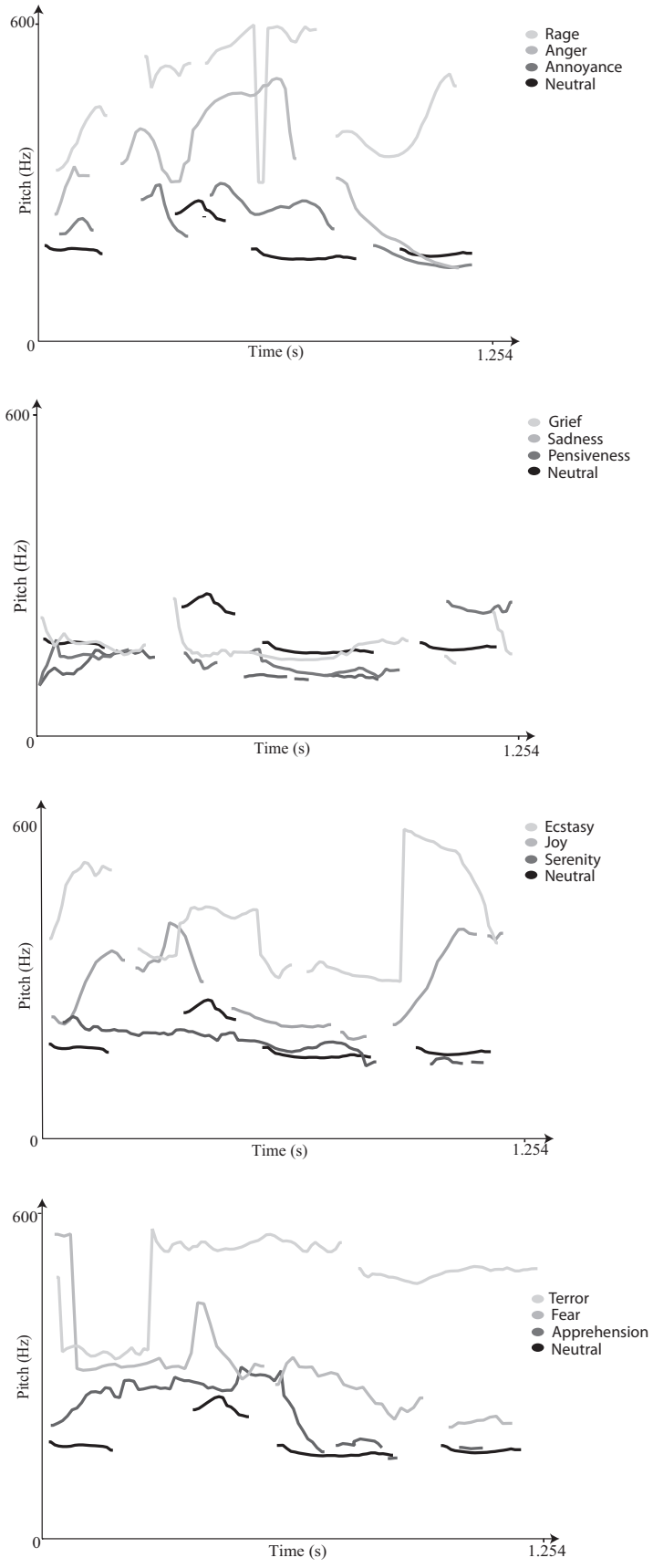

Fig. 3. F0-contours for four basic emotion groups and their intensities.

TABLE II
LIST OF EXTRACTED F0 FEATURES

Index	Feature
1	F0 Mean
2	F0 Median
3	F0 Standard Deviation
4	F0 Maximum
5	F0 Minimum
6	F0 Range
7	F0 Lower Quartile
8	F0 Upper Quartile
9	F0 Interquartile Range
10	F0 Kurtosis
11	F0 Skewness
12	F0 Slope
13	F0 Variation Rate
14	F0 Rising and Falling Ratio
15	Rising F0 Slope Maximum
16	Rising F0 Slope Minimum
17	Rising F0 Slope Mean
18	Falling F0 Slope Maximum
19	Falling F0 Slope Minimum
20	Falling F0 Slope Mean
22	F0 Rising Range Mean
23	F0 Falling Range Maximum
24	F0 Falling Range Mean

TABLE III
ACCURACY PERFORMANCE OF THE FIRST STEP OF CLASSIFICATION

Index	Group of emotion	AP
1	Anger	62.5%
2	Joy	40%
3	Fear	40%
4	Sadness	61,9%
5	Weighted Avg.	50%

D. Classification Results

In this studies classification process was divided into two parts. Firstly, all emotions were assigned to four groups representing primary emotions: anger, fear, sadness, and joy. Results of the first experiment are presented in Table III. Achieved results show that it is difficult to exactly recognize emotion basing only on F0 features even with such a small set of emotions. Best results were obtained, as well as in many other researchers, for anger.

Second step was classification of intensities inside each group. Results of this step are presented in Table IV. Both classifications were carried out using k-NN algorithm. For recognition of emotion intensities in a specific group accuracy performance greatly improves. One can observe some regularity for each group of emotions: best results were achieved for the weakest and strongest intensities, the worst results for primary emotions.

TABLE IV
ACCURACY PERFORMANCE OF THE SECOND STEP OF CLASSIFICATION

Group of emotion	Intensities	AP
Anger	Rage	83.3%
	Anger	42.9%
	Annoyance	71.4%
	Weighted Avg.	65%
Sadness	Grief	71.4%
	Sadness	42.9%
	Pensiveness	62.5%
	Weighted Avg.	59.1%
Joy	Ecstasy	60%
	Joy	47.4%
	Serenity	56.3%
	Weighted Avg.	54.5%
Fear	Terror	83.3%
	Fear	57.1%
	Apprehension	71.4%
	Weighted Avg.	70%

VI. CONCLUSIONS AND FUTURE WORK

In this paper, a new approach for recognizing emotions from speech signal has been proposed. The results of this investigation show that expression of emotion affects F0 contour. However, usage of features related solely to F0 does not provide satisfactory results. The average recognition accuracy of emotion group recognition is about 50%. For recognition of emotion intensities in a specific group accuracy performance greatly improves. One can observe some regularity for each group of emotions: best results were achieved for the weakest and strongest intensities, the worst results for primary emotions. Moreover, analysis of confusion matrix shows that if the classification is incorrect, results point at adjacent emotion of the same group.

Future work should concentrate on analyzing additional features, especially with regard to all vocal tract structures. Also other advanced classifiers should be tested. Furthermore samples for the training set should be recorded by professional actors, who are able to intentionally and properly express demanded emotions. This should widely improve results. Another and probably better solution would be a connection between audio, visual and semantic analysis system [16].

The issue of emotions recognition raises interest of people working in psychology, psychiatry, medicine and even in marketing area. Thus, intelligent, automatic system for emotion detection has a wide range of applications.

REFERENCES

[1] R. Plutchik, "The nature of emotion." *American Scientist*, vol. 89, July-August 2001.
[2] G. Iriea, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification." *IEEE Transactions on Multimedia*, vol. 12, October 2010.
[3] Y. Miyakoshi and S. Kato, "Facial emotion detection considering partial occlusion of face using bayesian network." *2011 IEEE Symposium on Computers and Informatics*.
[4] Z. Yang, "Multimodal datafusion for aggression detection in train compartments." February 2006.

[5] T. Kostoulas, T. Ganchev, and N. Fotakis, "Study on speaker-independent emotion recognition from speech on real-world data," *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction 2008*.
[6] M. Kotti, F.Patern'o, and C. Kotropoulos, "Speaker-independent negative emotion recognition," *2010 2nd International Workshop on Cognitive Information Processing*.
[7] J. Cichosz and K. Ślot, "Low-dimensional feature space derivation for emotion recognition." *ICSES 2006*.
[8] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," *Proc. ICASSP*, 2007.
[9] Z. Ciota, *Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy*, 2010, in Polish.
[10] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multi-lingual emotion recognition from speech - on language dependency of acoustic/prosodic features for anger recognition."
[11] N. Kamaruddin and A. Wahab, "Driver behavior analysis through speech emotion understanding." *Intelligent Vehicles Symposium*, 2010.
[12] R. Hidayati, I. Purnama, and M. Purnomo, "The extraction of acoustic features of infant cry for emotion detection based on pitch and formants." *Proc. Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, November 2010.
[13] E. Mower, M. Mataric´, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles." *Audio, Speech, and Language Processing*, vol. 19, July 2011.
[14] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers." *Proc. Eurospeech Lizbona*, 2005.
[15] K. Izdebski, *Emotions in the Human Voice Volume I Foundations*, 2007.
[16] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals." *Proc. IEEE Transactions on multimedia*, vol. 10, 2008.
[17] Y. Yeqing and T. Tao, "An new speech recognition method based on prosodic analysis and svm in zhuang language," *Proc. 2011 International Conference on Mechatronic Science, Electric Engineering and Computer*, 2011.
[18] A. Janicki and M. Turkot, "Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierajcych svm," *Proc. KSTiT*, 2008, in Polish.
[19] A. Shaukat and K. Chen, "Emotional state recognition from speech via soft-competition on different acoustic representations." *Proc. Neural Networks (IJCNN)*, 2011.
[20] A. Razak, R. Komiya, and M. Abidin, "Comparison between fuzzy and nn method for speech emotion recognition." *Proc. Information Technology and Applications, ICITA 2005*.
[21] T. Nwe, S. Foo, and L. D. Silva, "Detection of stress and emotion in speech using traditional and fft based log energy features." *Proc. ICICS-FCM 2003*.
[22] K. Soltani and R. Ainon, "Speech emotion detection based on neural networks." *Proc. Signal Processing and Its Applications 2007*.
[23] M. Gaurav, "Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech." *Proc. Spoken Language Technology Workshop 2008*.
[24] http://www.exaudios.com/.
[25] K. Scherer, R. Banse, and H. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures." *Journal of Cross-Cultural Psychology*, vol. 32, January 2001.
[26] T. Zieliński, *Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań.*, October 2009., in Polish.
[27] https://www.msu.edu/course/.
[28] S. Narayanan, C. Busso, and S. Lee, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection." *IEEE Transactions on audio, speech, and language processing*, 2009.
[29] C. Basztura, *Komputerowe systemy diagnostyki akustycznej.*, Wydawnictwo Naukowe PWN Warszawa 1996, in Polish.
[30] K. Ślot, *Rozpoznawanie biometryczne.*, December 2010, in Polish.