Speech Emotion Recognition System for Social Robots

Submitted: 15th January 2013; accepted: 4th June 2013

Łukasz Juszkiewicz

DOI: 10.14313/JAMRIS_4-2013/59

Abstract:

The paper presents a speech emotion recognition system for social robots. Emotions are recognised using global acoustic features of the speech. The system implements the speech parameters calculation, features extraction, features selection and classification. All these phases are described. The system was verified using the two emotional speech databases: Polish and German. Perspectives for using such system in the social robots are presented.

Keywords: *speech emotion recognition, prosody, machine learning, Emo-DB, intonation, social robot*

1. Introduction

Recently, a new field of robotics is being developed: a social robotics. A social robot is able to communicate with people in an interpersonal manner and achieve social and emotional goals [25]. Social robots are meant to collaborate with people and be their companions in applications such as education, entertainment, health care etc.

Emotions play a significant role in an interpersonal communication, so an ability to recognise other people's emotions from speech and adapting one's response is an important social skill. Person without it, would have difficulties with living in a society. Similarly, the social robot needs to be able to recognise expressions of emotions, so that it could communicate with man in the natural manner and could be his companion.

Information on emotion is encoded in all aspects of language, in what is said and in how it is said or pronounced, and the "how" is even more important than the "what" [29]. Considering all levels of language, from pragmatics down to the acoustic level, the following thing can be said: Starting with pragmatics, intention of speaker is highly correlated with his emotional state [22]. The literal meaning of an utterance is the most obvious display of emotions, so the statements or keywords such as "I am sad" can be treated as emotion indicators [19]. However, explicit expressions of emotions can be intended as ironic or not express true emotions of the speaker.

Another indicator of the emotions is a tone of a voice, that is the phonetic and acoustic properties of speech. For example, the cracking voice could be the evidence of an excitement. Voice quality and the prosody (pitch, intensity, speaking rate) have been best researched in the psychological studies. They also intuitively seem to be most important in expression of emotions. An often cited review of literature on emotions in speech was written by Murray and Arnott [17]. They refer to a number of studies which seem to have identified almost explicit correlation between emotions and acoustic parameters. However in the studies of different authors conflicting results can be found. This is probably due to the large variability of the expression of emotions and different variants of the certain emotions, such as "hot" and "cold" anger [8].

The system has been developed to recognise emotions in recorded statements on the basis of acoustic features of speech. Section 2 discusses the structure of the system and its implementation. Section 3 presents the results of system's verification. In conclusion the results are evaluated and further development plans are presented.

2. Speech Emotion Recognition System

Commonly used pattern recognition algorithms are applicable to the speech emotion recognition problem. However, there are at least two different approaches. One is estimating the short-time signal parameters and modelling their changes with the Hidden Markov Models or similar [16,18]. The other is extracting global features of the signal and applying statistical methods and various types of classifiers: SVM [5, 33], artificial neural networks [9, 30], decision trees [26] and other. The second approach was chosen — each utterance is analysed as a whole, so global features are extracted and then classified.

There is no compliance about an optimal set of features for speech emotion recognition. Moreover, such a set would depend on a number and type of emotions being recognised, language, recording conditions etc. Therefore, a standard approach is to extract very large features vector (even about 4,000) and then reduce the number of features to obtain a subset that has better discriminative power in particular task [23]. Commonly used features selection algorithms are principal component analysis, linear discriminant analysis, information gain ratio, sequential forward floating search [24,31,32].

The speech emotion recognition system [28] has the form of a set of MATLAB scripts using external tools — the programs Praat and Weka. Praat is a free (Open Source) program for phonetic analysis of speech. It can compute several parameters of speech: pitch, formants, spectrum, MFCC and many others [20]. Weka 3 (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning soft-

ware written in Java. It contains a collection of visualisation tools and algorithms for data preprocessing, filtration, clasterisation, classification, feature selection and regression modelling [10]. It is freely available under GNU General Public License. This form of the system allows easy modifications and facilitates testing various algorithms. The structure of the system is illustrated in figure 1. There are two phases of the system's operation: learning phase and evaluation phase. In the off-line learning phase feature selection and supervised learning of classifier is carried. In the evaluation phase the prepared system is used for on-line speech emotion classification.

2.1. Speech signal acquisition

The first stage of a recognition process is a speech signal acquisition — an acoustic signal has to be transformed into an electrical signal and then digitalised. Quality of a used microphone and a preamplifier plays important role — distortions may affect recognition accuracy. Regardless of the quality of the recording equipment there still is a problem of an environmental noise and the noise generated by motors and other moving parts of the robot itself. It was assumed that in recordings used for classification, signal to noise ration would be high enough and there would be no additional voices (one person speaking at the moment).

2.2. Speech signal parametrisation

To determine the emotion of the speech, six parameters of the speech signal were used: intensity, spectrogram, pitch, mel-frequency cepstral coefficients, long-term average spectrum and harmonics to noise ratio. Some of them were computed using several different methods, because they gave slightly different effects.

Intensity is the instantaneous sound pressure value measured in dB SPL, ie dB with respect to pressure of $2 \cdot 10^{-5}$Pa [20].

Spectrogram The speech signal is split into 16ms frames with a 10ms step. Each frame is Fourier transformed to compute the magnitude of the frequency spectrum. The phase is neglected. Then logarithm is taken from the magnitude spectrum, and the result is appended to a matrix. This matrix is called spectrogram [12].

Pitch is a fundamental frequency of the speech. It is produced by vocal folds (known commonly as vocal cords). Two algorithms were used to estimate the pitch: autocorrelation method [1] and cross-correlation method [21]. The pitch exists only for the voiced parts of the signal, so resulting waveform is discontinuous. Built in smoothing and interpolation functions are used to overcome this issue.

Mel-frequency Cepstral Coefficients (MFCC) are commonly used for parametrisation of the speech [20, 27]. Spectrum of the windowed signal is analysed by a bank of 26 bandpass filters with a central frequency equally spaced on a mel-scale, reflecting the subjectively perceived pitch. Subsequently, a discrete cosine transform is used to a logarithmised spectrum into a cepstrum. Only the first 12 coefficients are used. Additionally, a 13rd series is computed as the average of all 12 series of the coefficients.

Harmonics to noise ratio (HNR) is energy of the harmonic parts of the signal related to the energy of the noise parts. HNR is expressed in dB and computed using the autocorrelation method and the cross-correlation method [1].

Long-Term Average Spectrum (LTAS) is averaged logarithmic power spectral density of the voiced parts of the signal with an influence of the pitch corrected away [13].

2.3. Feature extraction

In order to extract more useful information from obtained parameters vectors additional vectors are derived from them:

- first and second order difference,
- values of local minima,
- values of local maxima,
- distance between adjacent extrema,
- value of difference of the adjacent extrema,
- slopes between the adjacent local extrema,
- absolute values of the two above.

The acoustic features are time-series — their length depends on the duration of analysed utterance. For classification purposes, it is necessary to convert the time series into a feature vector of fixed length. This is achieved by treating time series as outcomes of random variables and computing their statistics:

- arithmetic mean,
- median,
- standard deviation,
- global maximum,
- global minimum,
- first quartile,
- second quartile,
- range,
- interquartile range.

Further they will be referred to as the basic statistics.

Algorithm of extracting the features from the raw acoustic features is basically uniform (except for spectrogram and LTAS). For each of them, treated as time series, there are computed:

- basic statistics,
- linear regression coefficients,
- basic statistics of derived series,
- linear regression coefficients of local maxima,
- linear regression coefficients of local minima.

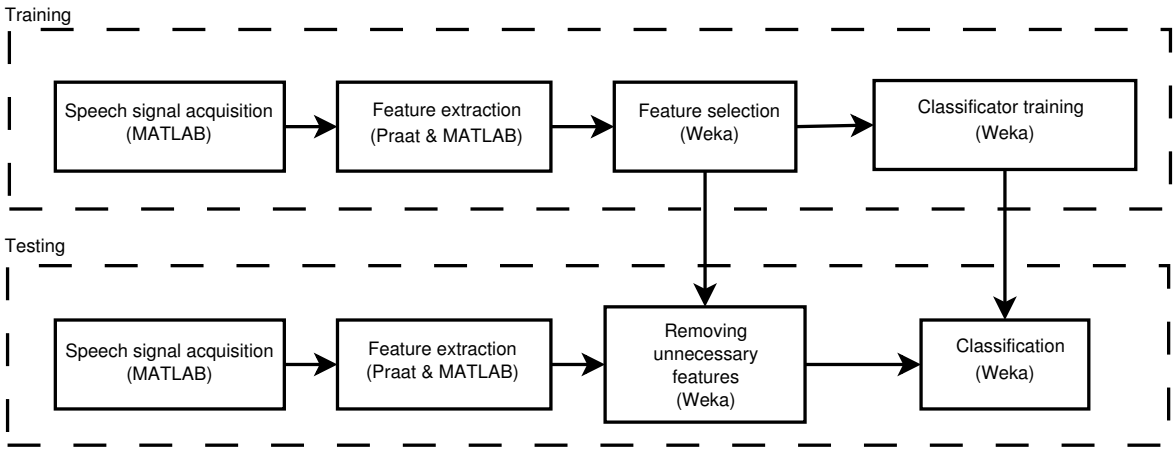

Fig. 1. Block diagram of speech emotion recognition system.

For each of the spectra forming spectrogram and for LTAS, there are computed:
- linear regression coefficients,
- centre of gravity, defined as $CG = \frac{\Sigma f_i E_i}{\Sigma E_i}$, where: f_i — i-th frequency, E_i — energy of f_i,
- 9th and 1st decile difference,
- slope between global maximum and minimum.

For LTAS, coefficients listed above are final features. For spectrogram basic statistics of them are computed to obtain final features.

Using this method 1722 features are generated. The structure of the feature vector is illustrated in figure 2. The number of features is then reduced in the feature selection process.

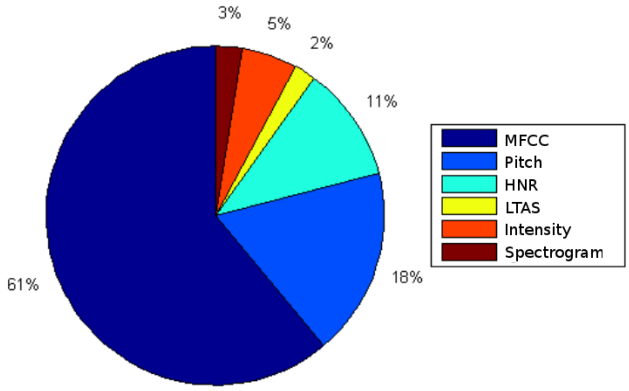

Fig. 2. Structure of feature vector.

2.4. Feature selection

Weka's function `AttributeSelection` was used to remove features that are redundant or not relevant for this particular task. This function requires subset evaluation and search functions. For evaluation correlation-based feature subset selection `CfsSubsetEval` was used — subsets of features that are highly correlated with the class while having low intercorrelation are preferred [11]. Search function was `BestFirst`, which searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

Feature selection is done in training phase. Therefore the selected feature vector depends on the training set.

2.5. Classification

Several different classifiers, provided by the Weka package, were tested:
- multilayer perceptron,
- support vector machine with sequential minimal optimisation,
- radial basis function network,
- Bayes network.

Results obtained from each of those classifiers were very similar. The feature selection stage seems to be crucial — selecting minimal vector of discriminative features makes the choice of classifier a minor problem. In section 3 the average recognition ratios for every classifier are presented. However, detailed classification results are presented only for `BayesNet` classifier — Bayes network that can use various search algorithms and quality measures [2]. Here, `SimpleEstimator` (estimating probabilities directly from data) and `K2` (implementing hill climbing) were used [7].

Advantage of Bayes Network classifier is the relatively low complexity while maintaining performance comparable with more complex classifiers. Additional advantage is explicit representation of knowledge — created network can be analysed or used for other purposes. Learning of Bayesian classifier is prone to mutual correlation of features — redundant ones could be dominant [14]. However, the algorithm used for feature selection removes redundant features, so this problem is eliminated.

3. Results

The tests were carried in two stages. In the first stage, two available databases of acted emotional speech: Berlin Database of Emotional Speech and Database of Polish Emotional Speech were used. They contain recordings registered in studio conditions, different from those that are expected in social robot

application. Those test, however, can verify the chosen methodology. The second stage is verification for the intended use of system — test recordings should meet typical responses to social robot and should not be recorded in such artificial conditions as mentioned databases.

3.1. Speech corpora

Berlin Database of Emotional Speech (Emo-DB) was recorded at the Technical University of Berlin [4]. Five male and five female carefully chosen speakers pretended six different emotions (anger, joy, sadness, fear, disgust and boredom) and neutral state in ten utterances—five consisting one phrase and five consisting two phrases. Every utterance is in German and has emotionally neutral meaning. After recordings 20 people were asked to classify emotion of each utterance and rate its naturalness. Utterances that were not classified correctly by more than four persons or considered as unnatural by more than two were discarded. Resulting database consists of 493 recordings.

Database of Polish Emotional Speech consists of 240 recordings [15]. Four male and four female actor speakers were asked to enact in five phrases five different emotions (anger, joy, sadness, fear and boredom) as well as the emotionally neutral state. Phrases are uttered in Polish and their meaning is emotionally neutral. The number of recordings is the same for each emotion. These recordings were evaluated by 50 humans — each of them was asked to classify 60 randomly selected samples. The average recognition ratio was 72%.

3.2. Speaker independent

Tests were carried out for the full set of recordings form both databases excluding those marked as disgust. There are no recordings of this emotion in Polish database, so it was ruled out to to keep the same conditions for both tests. Tenfolds cross-validation was used to evaluate classification accuracy.

Emo-DB A subset of Emo-DB used for experiments consisted of 489 utterances: 71 of joy, 62 of sadness, 69 of fear, 127 of anger, 81 of boredom and 79 neutral. Selected features vector was 123 elements long. Table 1 summarises numbers of the features derived from each of the parameters as well as the maximum and average information gain ratio (IGR) of those features. Despite the fact, that features derived from the MFCC and the pitch form a major part of vector, LTAS features have the higher average IGR. All other acoustic parameters are represented among selected features — each of the speech parameters is relevant and non-redundant. However, it should be noticed that HNR has significantly lower contribution than other parameters. System classified properly 396 (**81,98%**) instances using the Bayes net classifier. Table 2 shows confusion matrix and accuracy of classifications of each emotion. Results achieved using other classifiers were: 79.7% for perceptron, 80.1% for RBF network and 80.9% for SVM. In comparison in the study [30] accuracy of 79.47 % was achieved. In the studies [26] and [29] whole database (including disgust) was used and respectively 78.58% and 66,5% were archived.

Tab. 1. Number of features derived from the individual parameters of speech signal for Emo-DB

Parameter	No. of feat.	Max. IGR	Avg. IGR
MFCC	45	0.446	0.258
Pitch	29	0.434	0.241
Intensity	22	0.478	0.245
HNR	13	0.189	0.133
Spectrogram	8	0.429	0.213
LTAS	6	0.440	0.268
Σ	123		

Tab. 2. Confusion matrix for Emo-DB

		Classified as						
		N	J	S	F	A	B	
Actual class	N	68	0	1	3	0	7	86,1%
	J	2	37	0	8	24	0	52,1%
	S	3	0	58	0	0	1	93,5%
	F	1	8	4	52	4	0	75,4%
	A	0	11	0	4	112	0	88,2%
	B	8	0	3	1	0	69	85,2%

Polish Database In experiment for Polish language, all 240 recordings (40 of each emotion) were used. There were 86 selected features. Table 3 summarises numbers of the features derived from each of the parameters as well as the maximum and average information gain ratio (IGR) of those features. It can be noticed that the LTAS has significantly higher average IGR than the other parameters. System, using the Bayes net classifier, properly classified 177 (**73,75%**) instances (64,18% in [6]). Results achieved using other classifiers were: 71.7% for perceptron, 67.9% for RBF network and 70.8% for SVM. Table 4 shows confusion matrix and accuracy of classifications of each emotion.

Tab. 3. Number of features derived from the individual parameters of speech signal for Database of Polish Emotional Speech

Parameter	No. of feat.	Max. IGR	Avg. IGR
MFCC	41	0.467	0.287
Pitch	18	0.458	0.224
Intensity	12	0.390	0.278
HNR	6	0.320	0.223
LTAS	6	0.666	0.454
Spectrum	3	0.240	0.200
Σ	86		

3.3. Speaker dependent

Experiments with a speaker dependent recognition were carried out for the recordings from the

Tab. 4. Confusion matrix for Database of Polish Emotional Speech

		Classified as						
		N	J	S	F	A	B	
Actual class	N	31	0	2	3	0	4	77,5%
	J	2	32	0	0	6	0	80,0%
	S	0	0	35	1	0	4	87,5%
	F	1	0	8	21	8	2	52,5%
	A	0	7	0	4	29	0	72,5%
	B	3	0	7	1	0	29	72,5%

Database of Polish Emotional Speech. Feature selection and classification process was done separately for each speaker's 30 utterances. Length of the feature vectors and classification accuracy were summarised in table 5.

3.4. For social robots

In order to test the developed system for its usability for speech emotion recognition for the social robots, group of people was asked to utter phrases expressing four different emotional states: happiness, compassion, contempt and the neutral state. These statements were intended to meet the most common people's reactions to short-term contact social robot. The social robot Samuel (figure 3) was given as an example [3]. It was assumed that in such a task is better to recognise fewer emotions with the (expected) better accuracy. In comparison to the discussed databases, the meaning of phrases corresponds to the expressed emotion—it was concluded that in this case it is more likely. Example sentences are (English translation in brackets):

- Ale wyczesany robot! (What a cool robot!)
- Musi ci być smutno tak stać…(It must be sad to stand like this…)
- Weźcie to stąd! (Take this away!)
- Na podłodze leży dywan. (Carpet lies on the floor).

Recordings were taken in classroom in presence of many people at the same time, so conditions were far from the studio both in terms of acoustics and background noise. Electret microphone was used along with battery powered preamp and an external USB sound card. Sampling rate was set at 22050Hz, and the resolution was 16b. During experiment 160 sentences were recorded (40 for each emotion). One recording turned out to be corrupted, so it was ruled out.

Test set contained 159 recordings (39 for happiness, and 40 for each of other emotions). System performance was evaluated using Bayes net classifier and tenfold cross-validation. System correctly classified 110 instances (69%). Table 6 shows the confusion matrix and the accuracy of classification of each emotion. In the feature selection process 33 features were selected: 19 derived from MFCC, five from pitch, four from spectrogram, two from intensity, two from LTAS and one from HNR. Differently from previous experiments, the highest average IGR have features derived from the intensity (0.269) and those derived from LTAS have much lower avg. IGR (0.201). The lower avg. IGR have HNR (0.164).

4. Conclusions

This paper presents and discusses the speech emotion recognition system based on the acoustic features of speech, apart from its semantic. Verification of the system using the Berlin Database of Emotional Speech and Polish Database of Emotional Speech confirms the effectiveness of the chosen feature extraction, selection and classification methods. It should be noted, however, that in Polish language fear is clearly less recognised than other emotions, so is joy in the German language. For both languages, joy is often confused with anger — these emotions with opposite valence both have high arousal and are close to each other in acoustic feature space. This fact is important, because if the system was used by the social robot this type of mistake could result in incorrect response. One of the solutions could be weighting errors of misclassifing different pairs of emotions.

Verification for the target application showed that it is possible for the developed system to recognise emotions in recordings made in "non-sterile" conditions. Achieved recognition accuracy is promising for the future usage. However, social robot, designed to be human's companion, should be able to recognise more emotional states than the short-term contact robot. Therefore, further experiments and research should be carried out, especially concerning noise robustness.

Acknowledgement

This research was supported by Wroclaw University of Technology under a statutory grant.

Fig. 3. Social robot Samuel.

Tab. 5. Feature vector length and classification accuracy for speaker dependent tests.

Speaker	M1	M2	M3	M4	F1	F2	F3	F4
Length of feature vector	25	26	42	30	39	25	37	28
Classification accuracy	96,7%	100%	100%	100%	100%	96,7%	100%	100%

Tab. 6. Confusion matrix for social robot experiment.

		Classified as				
		Neu.	Cont.	Hap.	Comp.	
Cor. cl.	N	30	7	1	2	75%
	Ct	5	30	2	3	75%
	H	4	9	25	1	64%
	Cm	8	0	7	25	63%

AUTHOR

Łukasz Juszkiewicz* – Wrocław University of Technology, Institute of Computer Engineering, Control and Robotics, 50-370 Wrocław, Wybrzeże Wyspiańskiego 27, e-mail: lukasz.juszkiewicz@pwr.wroc.pl.

*Corresponding author

REFERENCES

[1] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", *Institute of Phonetic Sciences, University of Amsterdam, Proceedings*, vol. 17, 1993, pp. 97–110.

[2] R. R. Bouckaert, "Bayesian network classifiers in weka for version 3-5-7", *Network*, 2008.

[3] R. Budziński, J. Kędzierski, and B. Weselak. "Head of social robot Samuel – construction (in Polish)". In: *Prace Naukowe Politechniki Warszawskiej, Elektronika*, pp. 185–194, z. 175, t. I. Oficyna Wydawnicza Politechniki Warszawskiej, 2010.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. *A database of german emotional speech*, volume 2005, pp. 3–6. Citeseer, 2005.

[5] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms". In: *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, Washington, DC, USA, 2008, pp. 158–165.

[6] J. Cichosz and Ślot K., "Emotion recognition in speech signal using emotion-extracting binary decision trees". In: *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII): Doctoral Consortium*, 2007.

[7] G. F. Cooper and T. Dietterich, "A bayesian method for the induction of probabilistic networks from data". In: *Machine Learning*, 1992, pp. 309–347.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, no. 1, 2001, pp. 32–80.

[9] K. Dautenhahn. *Socially intelligent agents: creating relationships with computers and robots*, chapter Creating emotion recognition agents for speech signal. Multiagent systems, artificial societies, and simulated organizations. Kluwer Academic Publishers, 2002.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update", *SIGKDD Explor. Newsl.*, vol. 11, 2009, pp. 10–18.

[11] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, April 1999.

[12] S. Haykin, *Advances in spectrum analysis and array processing*, Prentice Hall advanced reference series: Engineering, Prentice Hall, 1995.

[13] E. Keller, "The analysis of voice quality in speech processing". In: *Lecture Notes in Computer Science*, 2005, pp. 54–73.

[14] P. Langley and S. Sage, "Induction of selective bayesian classifiers". In: *Conference on uncertainity in artificial intelligence*, 1994, pp. 399–406.

[15] Lodz University of Technology, Medical Electronics Division. "Database of Polish Emotional Speech". http://www.eletel.p.lodz.pl/bronakowski/med_catalog/docs/licence.txt.

[16] X. Mao, B. Zhang, and Y. Luo, "Speech emotion recognition based on a hybrid of HMM/ANN". In: *Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Informatics and Communications - Volume 7*, Stevens Point, Wisconsin, USA, 2007, pp. 367–370.

[17] I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of the Acoustic Society of America*, vol. 93, no. 2, 1993, p. 1097–1108.

[18] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models", *Speech Communication*, vol. 41, 2003, pp. 603–623.

[19] A. Osherenko and E. André, "Differentiated semantic analysis in lexical affect sensing", *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[20] D. W. Paul Boersma. "Praat: doing phonetics by computer (version 5.2.05)", 2010.

[21] S. Samad, A. Hussain, and L. K. Fah, "Pitch detection of speech signals using the cross-correlation technique". In: *TENCON 2000. Proceedings*, vol. 1, 2000, pp. 283 – 286.

[22] S. Schnall, "The pragmatics of emotion language", *Psychological Inquiry*, vol. 16, no. 1, 2005, pp. 28–31.

[23] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kossous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals". In: *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[24] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification". In: *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 864–867.

[25] B. Siciliano and O. Khatib, eds., *Springer Handbook of Robotics*, Springer, 2008.

[26] J. Sidorova, "Speech emotion recognition with TGI+.2 classifier". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2009, pp. 54–60.

[27] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch", *Journal of the Acoustical Society of America*, vol. 8, no. 3, 1937, pp. 185–190.

[28] Łukasz Juszkiewicz. *Postępy robotyki*, chapter Speech emotion recognition system for social robots (in Polish), pp. 695–704. Oficyna wydawnicza PW, 2012.

[29] T. Vogt. *Real-time automatic emotion recognition from speech*. PhD thesis, Technischen Fakultät der Universität Bielefeld, 2010.

[30] Z. Xiao, E. Dellandréa, W. Dou, and L. Chen. "Hierarchical Classification of Emotional Speech". Technical Report RR-LIRIS-2007-006, LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumiére Lyon 2/École Centrale de Lyon, March 2007.

[31] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "A hierarchical framework for speech emotion recognition". In: *Industrial Electronics, 2006 IEEE International Symposium on*, vol. 1, 2006, pp. 515–519.

[32] S. Zhang and Z. Zhao, "Feature selection filtering methods for emotion recognition in chinese speech signal". In: *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, 2008, pp. 1699–1702.

[33] J. Zhou, G. Wang, Y. Yang, and P. Chen, "Speech emotion recognition based on rough set and svm.". In: Y. Yao, Z. Shi, Y. Wang, and W. Kinsner, eds., *IEEE ICCI*, 2006, pp. 53–61.