Emotion recognition in speech signal using emotion-extracting binary decision trees

Jarosław Cichosz[1], Krzysztof Ślot[1]

Institute of Electronics, Technical University of Lodz,
Wolczanska 211/215, 90-924 Lodz, Poland. jarekcichosz@poczta.onet.pl, kslot@p.lodz.pl.

Abstract— The presented paper is concerned with emotion recognition based on speech signal. Two novel elements introduced in the method are an introduction of novel set of emotional speech descriptors and an application of a binary-tree based classifier, where consecutive emotions are extracted at each node, based on an assessment of feature triplets. The method has been verified using two databases of emotional speech on German and Polish, yielding very high recognition rates (72 %) for speaker-independent recognition.

Keywords — emotion recognition, speech analysis.

1 Introduction

Emotion recognition based on a speech signal is one of intensively studied research topics in the domains of human-computer interaction and affective computing. Recognition of emotional state is an increasingly important area in automated speech analysis due to several potential benefits that result from correct identification of subject's emotional state. Correct assessment of human emotion improves efficiency and friendliness of human-machine interfaces, allows for monitoring of psycho-physiological state of individuals in several demanding work environments, can be used to augment automated medical or forensic data analysis systems [1].

Despite a potential benefits and works that has been done in this field [2-7], the problem is still open [1], [8], [9]. One of the reasons of unsatisfactory performance of emotion recognition systems is a difficulty with an appropriate identification of a feature space to be used in classification. Although a lot of research has been done on defining good features of emotional speech signal, no widely acknowledged set of speech signal characteristics has been determined. Given some selected feature space, emotion recognition is typically performed using neural network , SVM, LDA, QDA classifiers [10], [11].

An objective of the research that has been reported in the following paper was to investigate possibilities of improving emotion recognition performance through derivation of novel descriptors of emotional speech and through an application of a binary-tree based recognition strategy. Emotional speech covering six different categories (joy, anger, boredom, sadness, fear and neutral) has been considered throughout our experiments. The proposed emotion recognition procedure involves a new category of speech signal features yielding very good recognition results (up to 76,3 % for speaker-dependent and up to 72 % for speaker independent cases).

A structure of the paper is the following. Emotional speech descriptors that have been introduced are characterized in Section 2. A structure of the adopted binary-tree classification strategy is presented in Section 3. Experimental evaluation of the proposed method, based on two different emotional speech databases – for German and Polish languages – is given in Section 4.

2 Emotional speech descriptors

Speech characteristics that are commonly used in emotion recognition can be grouped into three different categories. The first one includes frequency characteristics (for example a pitch and pitch-derived measures) which are related to voiced speech generation mechanism and vocal tract formation. The second group contains various energy descriptors that are related to speech production processes (such as mean or standard deviation of energy of an utterance). The third group comprises temporal features, which are related to behavioral speech production processes (such as utterance duration, pauses). Since the presented three groups of descriptors correspond to phenomena of different nature, one can assume that three-dimensional feature spaces can be efficiently used for the purpose of emotion characterization. This is one of the main premises that are used in our research.

A drawback of commonly used emotional speech descriptors is a lack of good information concerning temporal evolution of some particular parameter, which is selected for emotion quantification (e.g. mean pitch or mean energies are usually employed in recognition). To examine, whether information on evolution of a descriptor can improve recognition rates, we propose to include polynomial regression parameters into a feature set that characterizes emotional speech.

Another group of descriptors that has been explored throughout our research are energies in sub-bands of a speech signal spectrogram. Some of the considered emotions can correlate better with energies embedded in specific frequency intervals than with a total energy of a speech signal. Since our approach to emotion classification assumes extraction of a single emotion at every level of a binary tree, an adoption of such descriptors could prove beneficial.

2. 1 Regression parameters

The first group of newly proposed parameters includes linear and nonlinear regression parameters, estimated for some selected emotional speech descriptors. We assumed that a good representation of emotional load evolution throughout an utterance could be done either by linear regression (with parameters {a, b}) or by cubic regression (with coefficients{A, B, C, D}) of some speech characteristics s(t) (Fig 1.):

$$\{a,b\}:E[(s(t)-(at+b))^2]\to\min \tag{1}$$

$$\{A,B,C,D\}:E[\left(s(t)-(At^3+Bt^2+Ct+D)\right)^2]\to\min \tag{2}$$

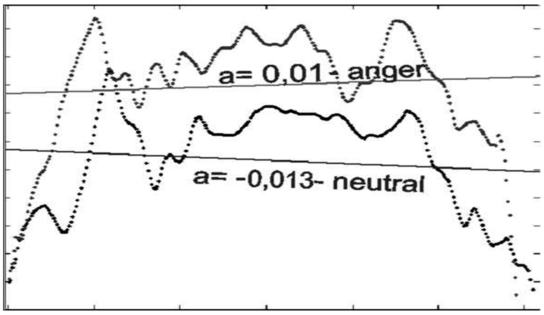

Fig. 1. Modeling of energy evolution with first-order regression for sentences uttered with various emotional loads.

Among many possible characteristics that can be approximated, we were particularly interested in modeling of the most successful emotional speech descriptors, such as e.g. pitch. For example, we considered approximations of smoothed pitch as well as approximations of pitch minima or maxima (Fig. 2).

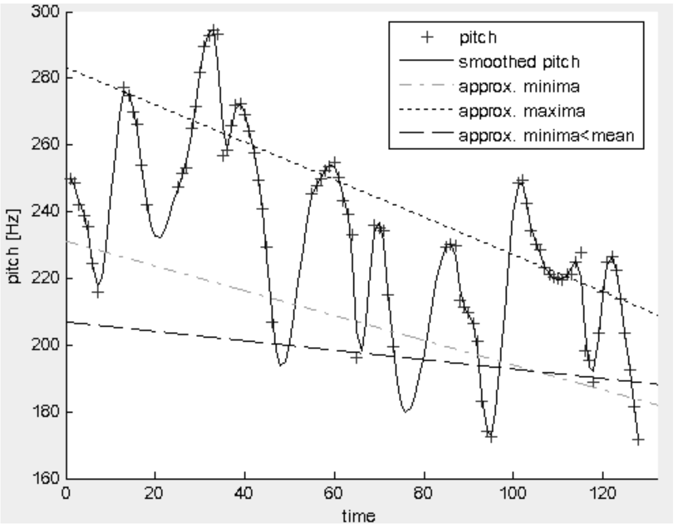

Fig. 2. Linear approximations of characteristic properties of pitch evolution

2. 2 Energy descriptors

The second group of parameters that have been introduced for emotion characterization are energy descriptors computed either within predefined spectrum sub-bands or within sub-bands defined adaptively with respect to mean value of a pitch (table I).

Table 1. Frequency bands used for computing speech signal energy

Band	Fixed bands [Hz]	Pitch-defined bands
E1	29 ~265	0~0,8
E2	265~1088	0,8~1,3
E3	1088~2205	1,3~1,5
E4	2205~3851	1,5~2,1
E5	3851~6056	2,1~4
E6	6056~8820	4~max

Example of sample frequency bands put on spectrogram have been shown in Fig. 3.

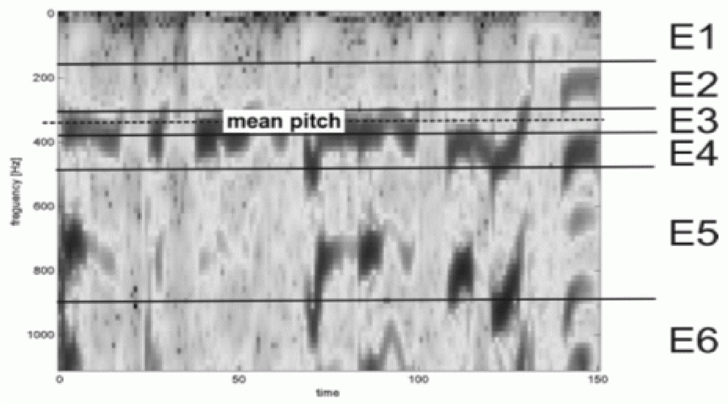

Fig. 3. Pitch- relative frequency bands that were used for derivation of energy descriptors.

A total pool of emotional speech descriptors that has been considered in our research comprised 102 elements, including both commonly used characteristics and the newly proposed elements.

3. Decision tree selection

Binary decision tree has been selected as a method for emotion classification. At each node of the tree a particular emotion gets identified, so that the tree has leaves at each of its levels (Fig. 4). A decision performed at each node (emotion “extraction”) is a two-class recognition problem, based on evaluation of a triplet of descriptors. An objective of the research was to find both the optimum structure of the tree and the optimum descriptor triplets. For a given tree structure, appropriate descriptors can be found using a basic feature selection procedure, where classification performance plays a role of a criterion function. Due to a limited number of emotions and an assumption of the three-element long feature vectors that are used at every node for decision-making, we were able to perform an exhaustive search procedure.

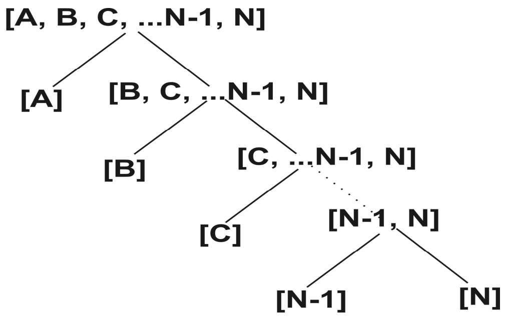

Fig. 4. Schematic diagram of the adopted decision-tree.

Feature selection procedure has been summarized in Fig. 5. Every candidate triplet was formed out of three feature sub-groups: frequency-related, energy-related and duration-related. Before applying such a triplet to the selection procedure, mutual correlation between descriptors was evaluated and, if it exceeded some threshold level, the corresponding triplet was dropped. A feature selection procedure was performed for all possible structures of a decision tree, i.e. for all permutations of the considered emotions (table 2).

Table 2. Emotion permutations tested in the feature selection process.

	Level I	Level II	Level III	Level IV	Level V	Level VI
1	neutral	anger	Fear	sadness	boredom	joy
2	neutral	anger	Fear	sadness	Joy	boredom
3	neutral l	anger	Fear	boredom	sadness	joy
4	neutral	anger	Fear	boredom	Joy	sadness
…	…	…	…	…	…	…
720	joy	neutral	Anger	fear	sadnes	boredom

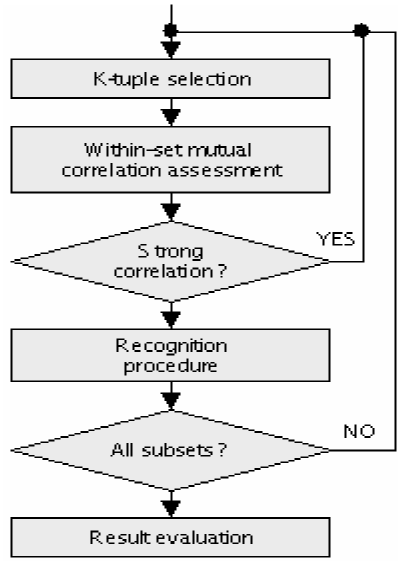

Fig. 5. Block diagram of feature selection procedure.

As a result of the presented procedure, a set of best performing decision trees associated with a set of best performing feature sets was identified.

4. Experimental method evaluation

The presented procedure has been verified using two different databases. The first one was a database of 240 sentences with six different categories of emotional load: **anger, fear, sadness, boredom, joy and no emotion** (a neutral emotion) [12], [13], [14]. The sentences were uttered in Polish by four actors and four actresses. The second database, "Berlin Database" [15], comprised 535 sentences uttered in German by five actors and five actresses. Utterance comprised seven different emotional categories **anger, fear, sadness, boredom, joy, disgust and no emotion (**we dropped disgust category to enable direct comparisons between both databases).

A performance of the proposed method in emotion recognition for both speaker-dependent and speaker-independent experiments has been evaluated. We assumed separate analysis of female and male recordings. Averaged results of the three best performing decision trees are shown in Table 3. Corresponding tree structures for trees that were performing the best for both languages are shown in Fig. 6.

Table 3. Averaged recognition results for the three consecutive, best performing decision trees.

Database	Best result	
	Speaker dependent	Speaker independent
Polish	**76,30**	**64,18**
German	**74,39**	**72,04**

As it can be seen, 76,30% (for Polish database) and 74,39% (for German database) recognition accuracy for speaker-dependent emotion recognition was obtained, whereas for speaker-independent case we got 64,18% for Polish and 72,04% for German. Results that have been obtained for speaker-independent case are slightly better, whereas results for speaker-independent case are significantly better than those given in [16].

Besides evaluating of an overall emotion recognition performance, we were interested in assessing a relevance of individual features. We computed frequencies of feature appearance among all triplets in the winning fifteen decision trees, for both Polish and German databases and for both speaker-dependent and speaker-independent cases. The results, presented in Fig. 7, identify a feature subset that appeared to be the most discriminative for the analyzed data. As it can be seen, five of the elements of the subset are the newly introduced ones.

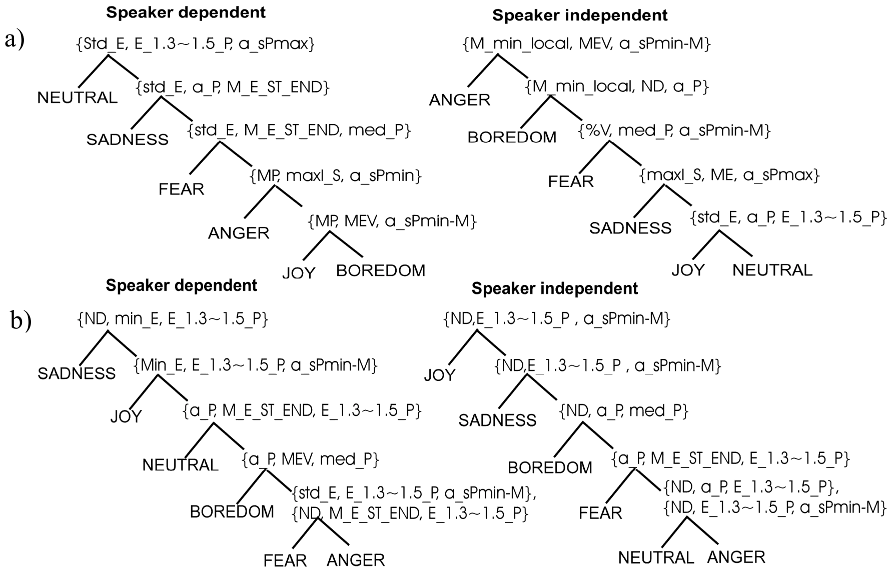

Fig. 6. The best performing decision-trees determined for emotion recognition in Polish database (a) and German database (b). Extracted emotions and corresponding best-performing feature triplets are shown at every node of the tree *(MP- mean pitch, maxI_S- magnitude of signal extreme value, std_E - standard deviation of energy, ND- normalized duration, M_min_local- mean value of local minima of pitch, a_P-regression coefficient for a pitch evolution, %V- voiced speech, ME-mean energy, b_E- linear regression coefficient for a energy evolution, M_E_ST_END- mean energy of all initial and final segments of voiced speech, MEV- mean energy of voiced speech, med._P- median pitch, min_P- minimum pitch, min_E minimum energy, E_1.3~1.5_P- energy in a frequency band- relative to a pitch (from1.3 to 1.5 of a mean pitch), min_sP- minimum smoothed pitch, a_sPmin-linear regression coefficient for local minima of smoothed pitch evolution, a_sPmax- linear regression coefficient for local mixima of smoothed pitch, a_sPmin-M- linear regression coefficient computed for these local pitch minima, which are located below its mean),*

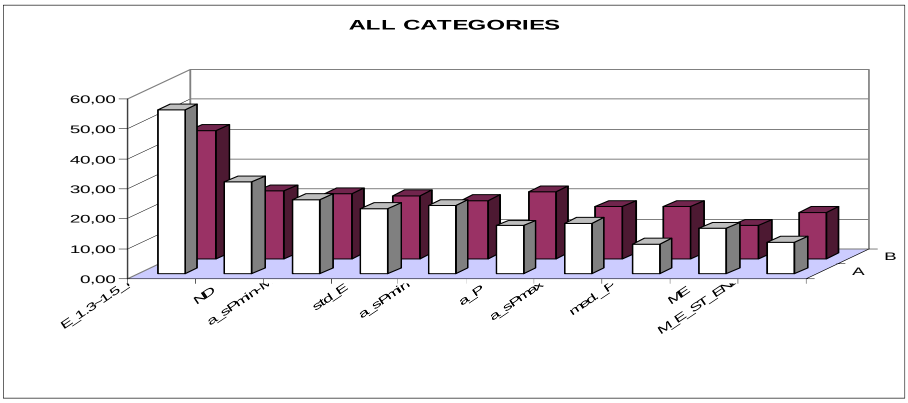

Fig. 7. An average percentage of appearance of the best performing features among the winning 15 trees for both databases (boxes in front – speaker-independent case; boxes in back – speaker dependent case).

5 Conclusion

The presented method of emotional speech classification through dichotomy-based decision-trees has been proved to produce very promising results. The experiments performed on two different databases, which comprised recordings of emotional speech, uttered in two different languages, yielded high recognition rates. This suggests that the proposed emotion-extraction approach can appear appropriate for the task realization.

We have also shown that the proposed, new emotional speech signal descriptors are very useful in emotion modeling. Among these descriptors, regression parameters of pitch and mean energy in low-frequency sub-bands consistently appeared to be among the best performing features. We believe that linear or nonlinear regression parameters are inherently good ways of emotion characterization, since they combine frequency and temporal description of the signal in a compact way.

Acknowledgments: This work supported by the Polish State Fund for Research Support in 2006-2008 under the Grant no.N51505831/2677.

References

[1] J.G. Taylor, K. Scherer, R. Cowie "Emotion and brain: Understanding emotions and modelling their recognition" *Neural Networks 18* pp. 313–316 (May 2005).

[2] R. Nakatsu, J. Nicholson, t. Naoko, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities". *Knowl.-Based Syst.* 13(7-8): 497-504 (2000)

[3] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms", *Speech Communication 40* (2003) 227-256.

[4] A. Batliner, R. Huber, J. Spilker, "The recognition of Emotion" Proc.of *Int. Conf. on Spoken Language Processing 2000.* pp. 122-130.

[5] R. Nakatsu, J. Nicholson, t. Naoko, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities". *Knowl.-Based Syst.* 13(7-8): 497-504 (2000)

[6] A. Batliner, R. Huber, J. Spilker, "The recognition of Emotion" Proc.of *Int. Conf. on Spoken Language Processing 2000.* pp. 122-130.

[7] R. Cowie, et al.: "Emotion recognition in human-computer interaction", *IEEE Signal Processing magazine*, vol. 18, no. 1, pp. 32-80, Jan. 2001.

[8] R. Cowie, E. Douglas-Cowie, M. Schroder, "Introduction", *ISCA Workshop*, Speech and Emotion, 2000

[9] E. Douglas-Cowie, R. Cowie, M. Schroder, "Speech and emotion", *Speech Communication*, Vol. 40, Issues 1-2, Pages 1-257 (April 2003).

[10] N. Fragopanagos, J.G. Taylor, "Emotion recognition in human–computer interaction", *Neural Networks*, Volume 18, Issue 4, Pages 389–405 (May 2005).

[11] Kwon O., Chan K., Hao J., Lee T. "Emotion Recognition by Speech Signals", *Proc. of Eurospeech 2003*, Genewa, p. 125-128, September 2003.

[12] J. Cichosz, K. Ślot, "Low-Dimensional Feature Space Derivation for Emotion Recognition"; *Proc. of Interspeech 2005*, pp. 477 – 480, Lizbon 2005.

[13] J. Cichosz, K. Ślot, "Application of selected speech-signal characteristics to emotion recognition in Polish language"; *Proc. of the ICSES'04*, p. 409 – 412, Poznań, Poland, 2004.

[14] J. Cichosz, K. Ślot, "Application of local signal descriptors to emotion detection in speech", *Proc. of the ICSES'06*, Lodz 2006.

[15] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. "A Database of German Emotional Speech"; Proc. of Interspeech 2005, pp. 1517 – 1520, Lizbon 2005.

[16] T. L. Nwe ,S. W. Foo, C. D. Silva, "Speech emotion recognition using hidden Markov models", *Speech Communication 41* (2003) 603 –623.