Speech Emotion Recognition

Ing. Matúš PLEVA, Ing. Lenka MACKOVÁ, Roman FRIČ

Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

matus.pleva@tuke.sk, lenka.mackova@tuke.sk, romano@klacno.sk

***Abstract*—The aim of this article is to present an introduction to the speech emotion recognition. The first tests of emotional speech recognition and the process of the emotional database recording are described, too. The purposes of emotion detection are depicted and possible algorithms of emotion detection from human speech communication are listed. The results of the first tests on the database are presented and compared with other institutions.**

***Keywords*—emotion recognition, Hidden Markov Models - HMM, nonverbal communication, short-time analysis**

I. Introduction

Nonverbal communication covers great area of different signals which follow human verbal communication. Nonverbal communication includes gestures, head and body movement, facial gaze, voice intonation and others.

Even when human emotions are hard to characterize and categorize the effort of its recognition increased in recent years due to the wide variety of applications that benefit from such technology.

Emotion recognition has lots of useful applications, for example in Human-Robotic Interfaces where robots can be taught to interact with humans and recognize human emotions (for example robotic pets could be able to understand to human commands), in call-centers where the „smart“ robotic system can replace human operators, in intelligent spoken tutoring systems to fill the gap between human and computer tutors and others.

Emotion recognition solutions depend on which emotion is wanted to be recognized by a machine and for what purpose. In general there are six basic emotional states: *neutral, happiness, fear, sadness, anger and disgust (or surprise).* In this article we focus in four emotional states: *neutral, happiness, sadness and anger.*

II. Speech Emotion Recognition

The main task of speech emotion recognition is appropriate voice processing. Human speech includes much more information than verbal text only. Each human voice is unique and in each human speech are coded emotions of the very speaker (different people use different tone in case of talking in anger, happiness or when even whisper). Besides human voice is changing, in cases of illness or even in the morning sounds, the human voice is a bit different than during the day.

Automatic emotion recognition of speech can be viewed as a pattern recognition problem [1]. The results of emotion recognition are characterized by: a) the features that are involved in the speaker's emotional state, b) the type of required emotions; c) the type of classifier used in the experiments and finally d) the database used for training and testing the classifier. The final results are obtained by comparison of classifiers in which case the same dataset and set of emotions is used.

Dellaert et al. [2] used in comparison three classifiers: the maximum likelihood Bayes classification, kernel regression, and k-nearest neighbor (K-NN) methods and was interested particularly in emotions of sadness, anger, happiness, and fear. The features used in this experiment were the pitch contour and the reached accuracy was 60%-65%.

In the experiment of Lee et al. [3] the linear discrimination, k-NN classifiers, and support vector machines (SVM) where used to distinguish two emotions states: negative and non-negative. In this case the maximum accuracy was 75%.

Navas et al. [4] provided two experiments of recognizing of joy, sadness, anger, disgust, surprise and fear. In the first one the short-term spectral features where used. The short-time statistics refer to the features obtained from the frame of the speech in short time intervals. The team of Navas used 18-MFCC and their first derivates. For each emotion in the database was built a GMM with 256 Gaussian mixtures.

In the second experiment long-term prosodic feature where used. The long-time statistics refer to the features calculated from speech parameters during a long time interval. The long-term features consist of different statistics calculated the pitch curve and it's first and seconds derivates, as well as from the first and second derivates of the power curve. The mentioned statistics where mean, variance, minimum range, skewness and kurtosis and finally jitter and shimmer values where also estimated and append to the final vector.

The results for emotional states differed approximately from 100% to 94%, in case of short-time analysis and from 97% to 83% in case of long-term analysis.

III. Emotional Database Recording

The recording of posed emotions in 16 neutral sentences – speech utterances (four different type of sentences) was done for only 30 speakers for now (male/female, different ages, accents), using 4 emotional states (angry, neutral, happiness, sad) and 4 background conditions (home, quiet, public place, office) using notebook as a recording device.

The four different neutral sentences without an emotional meaning were spoken using four emotional speech states.

Recordings were done using notebook microphone in 48kHz 16bit WAV PCM format. Every speaker records 16 sentences in one background environment (home, quiet, public place or office). Then the recordings were converted using sox [5] tool to RAW format recordings. Next the recordings were labeled according to the emotional state posed in the speech segment.

After the Master Label File (MLF) [6] was generated using automated Perl scripts from the whole database, the configuration files for HTK tools were prepared (grammar file, dictionary, list of all speech files, list of emotions, HMM prototype, etc.).

IV. Feature Extraction and Training

Only a short-time analysis [4] was done using the HMM models. The MFCC (Mel-Frequency Cepstral Coefficient) coefficients were used for feature extraction from the speech segments. More feature extraction configurations were tested. The energy (0), log-energy (E), delta (D), delta-delta - acceleration (A) coefficients and cepstral mean subtraction (Z) parameter were tested and the recognition results were compared. The best configuration was when computing thirteen MFCC log-energy, delta and acceleration coefficients with cepstral mean subtraction (E_D_A_Z) as we can see on the Figure 1 below.

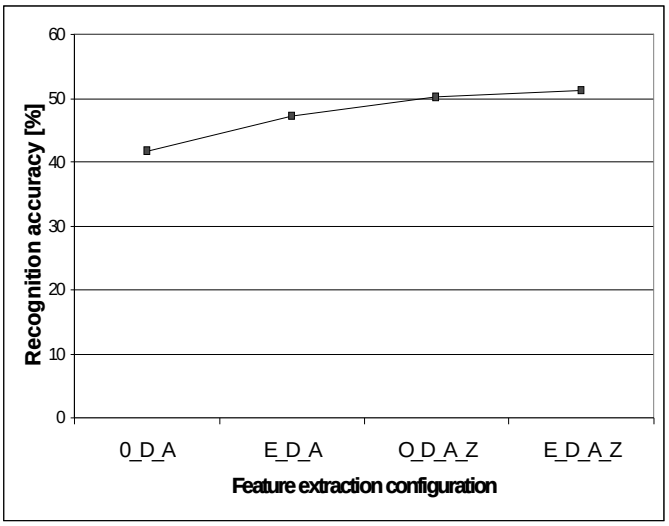

Fig. 1. The emotional states of the spoken segment recognition accuracy changed according to the feature extraction configuration.

The next step was testing different types of HMM prototypes (Fig 2). The left-right, and different ergodic prototypes from 3 to 7 states were trained and than tested using the same feature extraction configuration. The best result was reached using 5-state ergodic model with a small possibility to transit between all states (except the entering and emitting state).

Also a testing of different numbers of the PDF (Power Density Function) mixtures was done using from 2 to 256 PDF on the state. The recognition accuracy raised together with the number of the PDF mixtures on the state (Fig. 3). So for the first recognizer configuration the 256 PDF mixtures on the state of the HMM model was used.

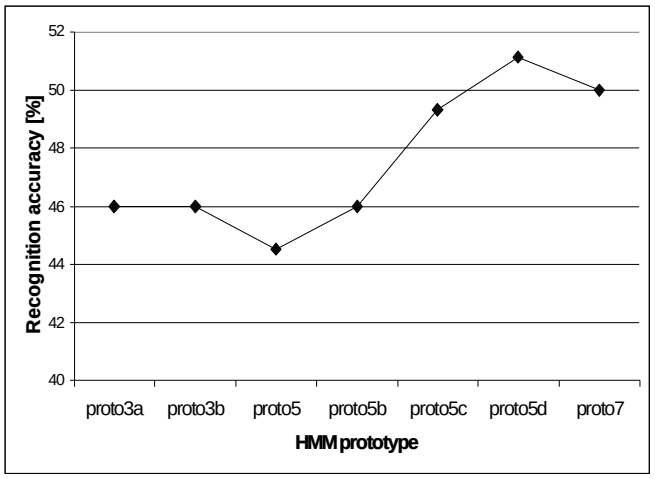

Fig. 2. The emotional states of the spoken segment recognition accuracy changed according to configuration of the HMM prototype and the number of the states.

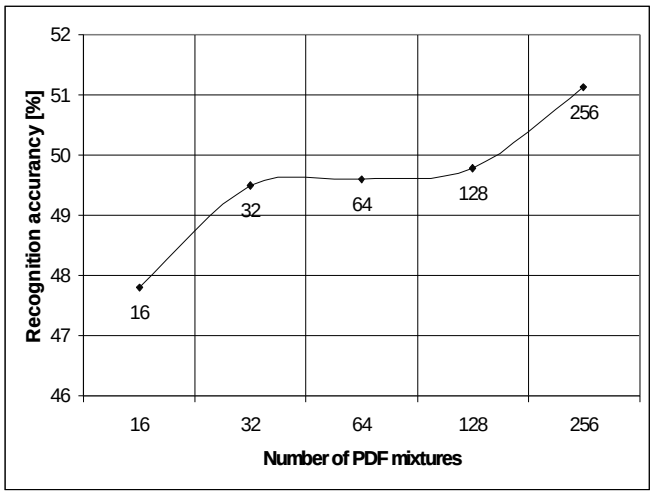

Fig. 3. The emotional states of the spoken segment recognition accuracy raised according to the number of PDF mixtures on the HMM model state.

V. Testing

The testing was realized using Viterbi decoding algorithm from HTK Tools (HVite). The recorded speech utterances were divided to training and testing part in the rate of 2:1. The training part was used only for training and the testing part was used only for automatic recognition. The HMM model produced sometimes more than one result for one speech utterance. For example sadness was recognized for the first half and neutral for the second half of the speech utterance sentence. That's why the sum of the recognized emotions in the anger column is more than one hundred percent. The complete confusion matrix is shown below in the Table 1.

TABLE I
Confusion Matrix [% of Recognized Emotion to Original]

Recognized:	*Original Emotion*			
	Anger	*Sadness*	*Happiness*	*Neutral*
Anger	*66*	*8*	*17*	*6*
Sadness	*5*	*50*	*11*	*33*
Happiness	*26*	*11*	*54*	*8*
Neutral	*14*	*34*	*11*	*40*

As we can see the anger and the happiness emotional speech is better recognized. The problem is to recognize the difference between neutral and sadness emotional speech. Both are uncertain also for humans, because there are differences not only between the men and women expression [7] of the emotions, but also there are differences between emotion expressions in the various cultures.

Next the verification of the results using emotional speech utterances recognized by humans will be done. All speech utterances will be recognized by humans and only utterances where the human recognized emotion corresponds with the original emotion will be used for training the automatic emotional speech recognition engine.

Then also testing will be done on the filtered utterances of the posed emotional speech. This process will filter the error, which corresponds to a bad acting of the emotions of the non-professional speakers in the recorded database. After the testing will be done the comparing of the results will answer the question how important is a good acting in posed emotional database recording.

VI. Comparing The Results

So finally only a short-time analysis [4] using the 5-state ergodic HMM models was done with overall 51.13% accuracy using MFCC_E_D_A_Z parameterization, 256 PDF on each of the 5-state ergodic HMM prototype. In the future will this emotional speech recognition algorithm trained and tested using a bigger spontaneous emotional database with also the background conditions annotated (using F-measures [8]) after the annotation process will be done from the previously recorded live TV discussions.

This non-verbal information about the emotion of the spoken speech segment could give the hearing impaired users of the recognized speech utterances database a better understanding of the automatic transcribed texts, which could lead to better information efficiency [11].

The 51.13% recognition accuracy for recognizing one of four possible emotional states is not clearly comparable with the tests of the other institutions mentioned before. For example Dellaert et al. [2] reached 60% overall accuracy but recognizing 6 emotional states in speech utterances. Navas et al. [4] from Basque university reached 95% overall accuracy for recognizing six emotions, but on various bigger emotional speech databases as Idoia (665 emotional speech utterances) [9], Karolina and Pello (700 sentences) [10].

VII. Conclusion

The interesting application for speech utterance emotion recognition is the automatic transcription of the speech with some emotional text labels. This additional information could give the user of the transcribed texts database the additional information about the meaning of the sentence. This additional emotional state information is useful when searching some special meaning of the words or presenting the texts without the audio data for hearing impaired or when the audio channel is not available.

Also a very useful application for emotional speech recognition engine is the area of the human machine interaction and virtual agents. There are some experiments with virtual agents, interacting on the emotional state of the human and trying to also place the speech synthesis [12] in the dialog and the whole virtual agent behavior in the same emotion, and make the communication friendlier for the end-users.

These preliminary results give us the knowledge for building the spontaneous emotional database, and spontaneous emotional speech recognition engine.

In the future we would like to tune the algorithms of training and the feature extraction process. Also the posed emotional speech database is not useful for spontaneous emotional speech recognition, so the annotated spontaneous speech database will be built from previously recorded live TV discussions.

Acknowledgment

Research described in the paper was supported by the Slovak Ministry of Education under research projects AV 4/0006/07, AV 4/2016/08, VEGA 1/4054/07 and MVTS COST2102/07.

References

[1] S. Yacoub, S. Simske, X. n Lin, and J. Burns, "Recognition of Emotions in Interactive Voice Response Systems**, "** HP Laboratories Palo Alto, 2003, pp.2

[2] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," in *Proc. of ICSLP 1996*, Philadelphia, PA, pp. 1970 -1973, 1996

[3] C. Lee, S. Narayanan, and R. Pieraccini, "Classifying Emotions in Human-Machine Spoken Dialogs", in *Proc .of International Conference on Multimedia and Expo*, Lausanne, Switzerland, August 2002

[4] E. Navas, I. Hernáez, I. Luengo, I. Sainz, I. Saratxaga, and J. Sanchez, "Meaningful Parameters in Emotion Characterisation," In: *Verbal and Nonverbal Communication Behaviours*, Lecture Notes in Computer Science 4775, Springer Verlag 2007, pp. 74 - 84, Editors: A. Esposito et al., Revised selected and invited papers from COST Action 2102 international workshop in Vietri sul Mare, Italy, March 2007

[5] http://sox.sourceforge.net/ - cached March 2008

[6] S. Young, "ATK: An application Toolkit for HTK", version 1.3, Cambridge University, January 2004

[7] T. Vogt, and E, André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conference (LREC 2006)*, pp. 1123 - 1126

[8] J. Žibert, F. Mihelič, J.P. Martens, H. Meinedo, J. Neto, L. Docio, C. Garcia-Mateo, P. David, J. Zdansky, M. Pleva, A. Čižmár, A. Žgank, Z. Kačič, C. Teleki, and K. Visci, "COST278 broadcast news segmentation and speaker clustering evaluation", In: Interspeech, Lisboa, 2005. Bonn, Universität Bonn, 2005. p. 629-632. ISSN 1018-4074

[9] E. Navas, A. Castelruiz, I. Luengo,, J. Sanchez , and I. Hernáez, "Designing and recording an audiovisual database of emotional speech in Basque," in Proc. of *Language Resources and Evaluation Conference (LREC 2004)*, pp. 1387 - 1390

[10] I. Saratxaga, E. Navas, I. Hernáez,and I. Luengo, "Designing and recording emotional speech database for corpus based synthesis in Basque," in Proc. of *Language Resources and Evaluation Conference (LREC 2006)*, pp. 2127 - 2129

[11] M. Pleva, A. Čižmár, J. Juhár, S. Ondáš, and M. Mirilovič, "Towards Slovak Broadcast News Automatic Recording and Transcribing Service", In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Selected papers from COST Action 2102 International Workshop. Volume Editor(s): A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis. Lecture Notes in Computer Science, to be published in Springer Verlag 2008, pp. 165-176

[12] J, Pribil, and A. Pribilova, "Emotional style conversation in the TTS system with cepstral description," In: *Verbal and Nonverbal Communication Behaviours*, Lecture Notes in Computer Science 4775, Springer Verlag 2007, pp. 65 - 73, Editors: A. Esposito et al., Revised selected and invited papers from COST Action 2102 international workshop in Vietri sul Mare, Italy, March 2007