A STUDY OF THE EFFECT OF EMOTIONAL STATE UPON TEXT-INDEPENDENT SPEAKER IDENTIFICATION

*Marius Vasile Ghiurcau** , *Corneliu Rusu*†

Technical University of Cluj-Napoca
Faculty of Electronics, Telecom. and Inf. Tech.
Signal Processing Group
Cluj-Napoca, Romania

Jaakko Astola

Tampere International Center
for Signal Processing,
Tampere University of Technology
Tampere, Finland

ABSTRACT

In this paper we evaluate the effect of the emotional state of a speaker when text-independent speaker identification is performed. The spectral features used for speaker recognition are the Mel-frequency cepstral coefficients, while for the training of the speaker models and testing the system the Gaussian Mixture Models are employed. The tests are performed on the Berlin emotional speech database which contains 10 different speakers recorded in different emotional situations: happy, angry, fear, bored, sad and neutral. The results show an important influence of the emotional state upon text-independent speaker identification. In the end we try to give a possible solution to this problem.

***Index Terms*—** speech signal processing, text-independent speaker recognition, emotions, MFCC, GMM

1. INTRODUCTION

The problem of speaker recognition has been addressed many times in the open literature and several solutions have been proposed. Gaussian Mixture Models (GMMs) represent a tool widely used in both speech and speaker recognition. Moreover, most of the baseline systems for text-independent speaker recognition use Mel-frequency cepstral coefficients (MFCCs) in a GMM framework. Various enhanced versions of the GMMs have been proposed and in general the results show a very high efficiency of the GMMs in speaker recognition.

It is well known that speaker identification has various applications in real life. Most of them belong to the different kinds of security systems (building access systems), where the human voice serves as a key. In this paper our work focuses on the problem of text-independent speaker identification, but in a special situation, when the recorded speakers present different emotional states. We all know, human beings are not machines and quite frequently are being overwhelmed by emotions: happiness, fear, anger, sadness, boredom and so on. These emotions are present in our everyday life. The studies suggest that approximately 10% of the human life is unemotional [1, 2] while the rest is affected by various emotions. Most of the times, one can not really control these emotional states. Consequently, there is a question that arises: do the speaker verification/identification systems perform well when the speakers have different emotional states?

Recognition of human emotional state is a topic under development in the last years. Among the possible applications suggested in the literature one can point out educational software [3], telephone response services (call centers [4]), learning environment or entertaiment [5]. So far, most of the work conducted in the 'emotional field' implied emotion recognition. Giannakopoulous in [6] suggests a method for extracting affective information using speech data from movies. In [7] it is presented a regression approach to music emotion recognition, with applications in music retrieval field. Another interesting method for recognizing emotions in speech can be found in [2]. Most of the work in the area of emotion recognition use either audio or visual informations. Even though, there are some studies which suggest a combined system with both audio and visual features [5] for enhancing the results of the emotion recognition systems.

As to our knowledge very few work has been done for speaker recognition in emotional environment. Shahin in [4] tried two different approaches for text-dependent speaker recognition in emotional environment with performances between 50% and 60% depending on the emotional state. However, these tests were performed separately, depending on the gender of the speaker.

In the present study we propose a comparison between the performances of a classical text-independent speaker identification system, trained and tested with speakers recorded in neutral states, and then, trained and tested with speakers that simulate different emotional states. At the end, we propose a solution for increasing the performances of such a speaker identification system and also suggest some other future work possibilities.

*This research was supported by PRODOC POSDRU/6/1.5/S/5 ID 7676.
†This research was supported by CNCSIS Grant ID 162/2008.

The rest of the paper is organized as follows. At the beginning some aspects about the theoretical background of this paper are presented in Section 2. Practical work is provided in Section 3. Section 4 presents the results of our research and a short discussion. Some conclusions and future work possibilities can be found in Section 5.

2. GAUSSIAN MIXTURE MODELS

A Gaussian Mixture Model can be written as a weighted sum of M component densities and has the following form [8]:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x)$$

where x is a d-dimensional random vector, $p_i(x), i = 1, ..., M$, is the component density and $w_i, i = 1, ..., M$, is the mixture weight.

The component densities are d-variate Gaussian functions given by [8]:

$$p_i(x) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma_i)}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]$$

where μ_i is the mean, Σ_i is the covariance matrix, d is the number of features incorporated into every feature vector.

The weights w_i have to satisfy the following relation: $\sum_{i=1}^{M} w_i = 1$. Each model can be written as a function of the following parameters: $\lambda = (w_i, \mu_i, \Sigma_i)$, $i = 1, ..., M$.

The log of the likelihood function is [9]:

$$\ln\ p(x|w, \mu, \Sigma) = \sum_{n=1}^{N} \ln\left\{\sum_{i=1}^{M} w_i p\left(x_n|\mu_i, \Sigma_i\right)\right\}$$

where $x = x_1, ..., x_N$

Finally, for finding the maximum likelihood solutions for the models, different algorithms are used. In the present study, the expectation-maximization (EM) algorithm was employed.

3. PRACTICAL WORK

3.1. Database description

The speech database used in this study is the Berlin Database of Emotional Speech [10] recorded at the Technical University of Berlin. The speakers are represented by 10 actors (five males and five females) each of them simulating different emotional states when asked to recite 10 different German utterances (five short and five longer sentences) [11]. Overall the database contains more than 500 different utterances labeled with the following emotional states: neutral, anger, fear, joy (happiness), sadness and boredom. In order to be as close to reality as possible the ten 'actors' were selected from a group of 40 people that were invited to a preselection session where the judging was performed taking into account the naturalness and recognizability of the performance [11].

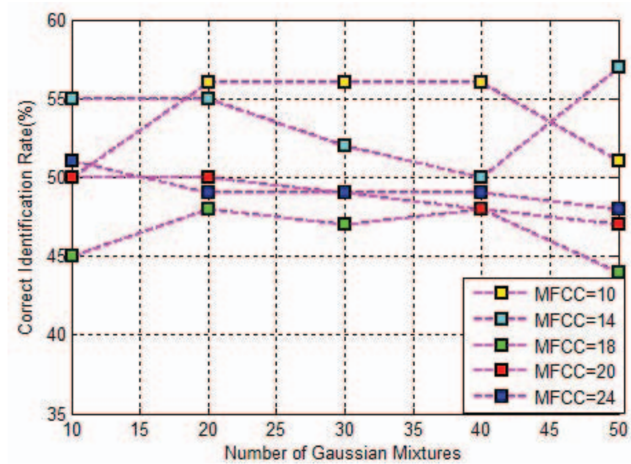

Fig. 1. Results for text-independent speaker identification when training is done with neutral utterances and various emotional states are used for testing

3.2. Experimental setup and simulations results

At the beginning, the Mel-frequency cepstral coefficients were extracted from the signals. Then the training of the models and testing using the Gaussian-Mixture Models was employed. For each of the speakers there were available around 10 utterances of a few seconds of speech. All the recordings were sampled at 16 kHz (16 bits mono *.wav files). For the pre-emphasis an FIR filter with the pre-emphasing coefficient $a = 0.97$ was used. The signal was divided in 256-sample frames with an overlap of 128 samples (50% overlap) and the MFCCs were computed for each frame.

In our experiments we have tried using 10 to 24 cepstral coefficients and the results were compared. First coefficient was all the time discarded, because it is dependent of the channel gain. Along with MFCCs we have also used the delta coefficients (first order differences of the MFCCs).

Four main experiments were performed:

Experiment 1: We took ten different utterances of the same emotional state of the speakers. 9 of the utterances were used for training and 1 for testing. We have repeated this procedure 10 times, every time choosing another test utterance (leave one out cross validation method). At the end, we computed the average of our results. Overall, correct identification rates between 99% and 100% were achieved (for both neutral and bored state maximum identification rate was encountered).

Experiment 2: We have repeated *Experiment 1*, just that we have used neutral utterances for the training phase and recordings with different emotional states for the testing

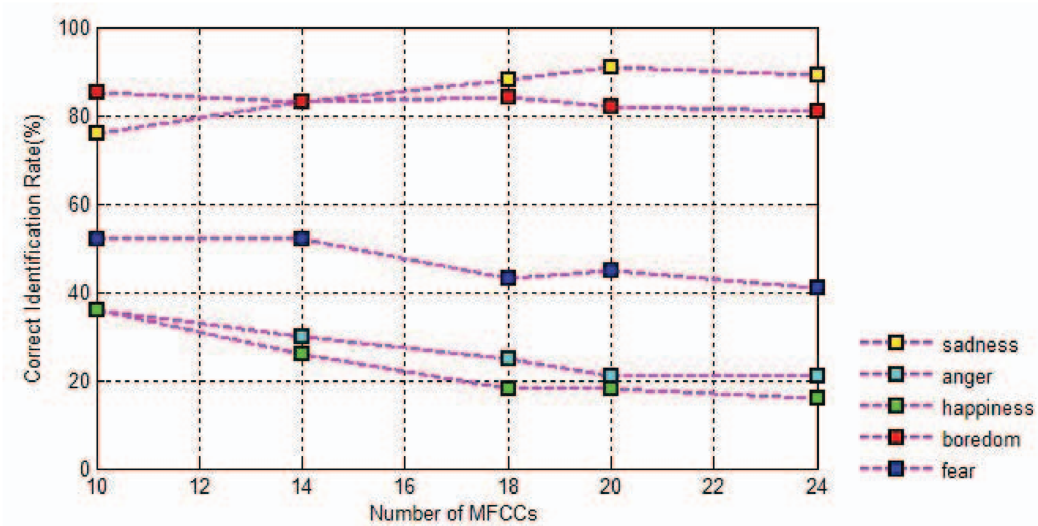

Fig. 2. Comparisson of the results for text-independent speaker identification when training is done with neutral utterances and testing is done using in each case a different emotional state

phase (for each speaker, we had ten test recordings, 2-3 recordings of each remaining emotional state). Figure 1 presents the overall correct identification rates for different combinations of the number of MFCCs and number of Gaussian mixtures.

Experiment 3: We wanted to find which are the emotions that influence mostly the text-independent speaker identification system. To this end, we trained again the system with neutral utterances and tested it using each time different emotional states; first the 'happy' instances, then angry, fear, bored and finally the sad states.

We noticed that for this particular situation the best results are obtained mostly when using 50 Gaussian mixtures and different number of MFCCs, so we decided to present the results only for this cases, due also to the high amount of data. Figure 2 presents the results for 10, 14, 18, 20 and 24 MFCC coefficients for all of the emotional states.

Experiment 4: The poor results from the previous experiments had to be improved, consequently some modifications had to be imposed. For each of the 10 speakers, 35 utterances were arbitrarily selected, each of the groups containing 5-6 recordings of every emotional state available. We took 34 of the recordings from each speaker and trained a GMM model. We tested the system using the remaining recording. The process was repeated 35 times. The overall results are presented in Figure 3.

In our experiments we also tried increasing the number of MFCC coefficients and/or the number of Gaussian mixtures but there was no improve in the final results. We also tried using only the MFCC coefficients, without the delta coefficients, but the outcomes decreased.

4. RESULTS AND DISCUSSIONS

As it was expected when the same type of emotional state is used for both training and testing the results are very good, with values between 99% and 100%. Achieving a maximum correct identification rate it is quite normal taking also into account the small number of speakers.

When neutral state is used in training and the rest of the emotional states are used for testing the system fails. None of the tested MFCC/GMM combinations managed to get scores above 60% which could be assimilated to a total fail in real applications.

Next goal was to find how each of the emotional states influence the identification process. It seems that anger and happiness represent the most difficult situations; for both of them we received correct identification rates that vary between 16% and 36% depending on the number of MFCCs. For fear, in the best case we can reach to 52% correct identification rate. The best results are obtained for boredom and sadness, when we can reach up to around 90% in some situations. Even though, the results can not be considered as satisfying.

It seems that if we train the system with all the available emotional states and also test it with different emotional states the results increase significantly from the previous cases, to scores of around 98%. The best score of 98.57% was achieved for 16 or 20 MFCC coefficients and 40 and respectively 50 Gaussian mixtures.

5. CONCLUSIONS AND FUTURE WORK

As it was expected, MFCC and GMM perform very well in text-independent speaker verification. When emotions alter the human voice, the performances of the speaker verifica-

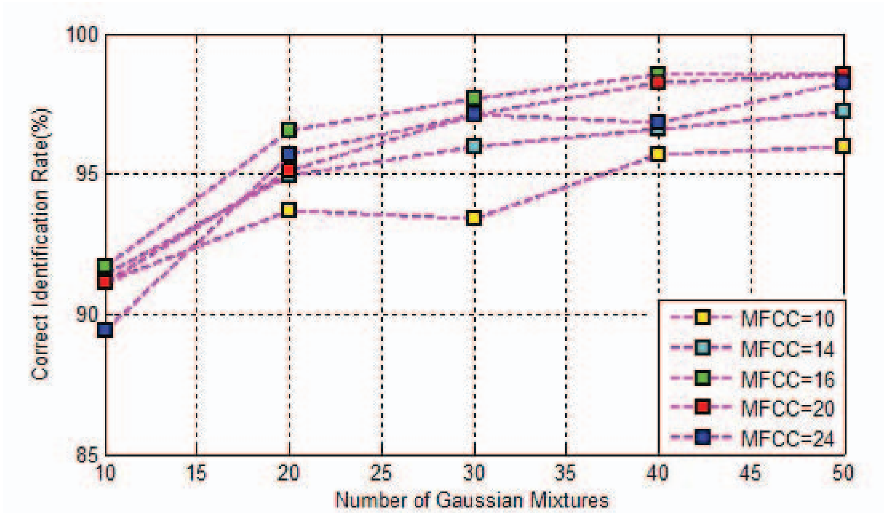

Fig. 3. Text-independent speaker identification when both training and testing is performed using various emotional states

tion systems decrease significantly. If training of the system is done using utterances of the speakers in different emotional states, the correct classification rates increase up to approximately 98%. As a consequence, for the future it would be a good idea to train a speaker identification/verification system with utterances in different emotional states. Even though the increase is considerable in this case, it is still not sufficient. In order to be a viable solution, the rates should hold up to over 99%. Consequently, future work is needed in this area.

A possible solution could be using a 'two-step' identification system, in which firstly the emotion is identified, and furthermore we proceed to speaker identification using trained models of the particular identified emotion. For improvement of the existing system, Support Vector Machines ca be a solution, also a combination of GMM and SVM should be tried. Using MFCC combined with other features, such as fundamental frequency, may improve the results, even though fundamental frequency is a prosodic speech feature strongly dependent of the emotional states.

6. REFERENCES

[1] The HUMAINE Portal, "Research on emotions and human-machine interaction," `http://emotion-research.net`.

[2] Z. Xiao, E. Dellandrea, L. Chen, and W. Dou, "Recognition of emotions in speech by a hierarchical approach," in *Proc. IEEE ACII 2009*, Amsterdam, Netherland, 2009.

[3] R.W. Picard, *Affective Computing*, Cambridge, MIT: Media Laboratory - Perceptual Computing Section - Technical Report no. 321, 1995.

[4] I. Shahin, "Speaker recognition systems in the emotional environment," in *Proc. IEEE ICTTA '08*, Damascus, Syria, 2008.

[5] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," in *IEEE Transactions on Multimedia,*, vol. 10, no. 4, pp. 659-668, 2008.

[6] T. Giannakopoulos, A. Pikrakis, and S. Theoridis, "A dimensional approach to emotion recognition of speech from movies," in *Proc. IEEE ICASSP '09*, Taipei, Taiwan, 2009.

[7] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen, "A regression approach to music emotion recognition," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448-457, 2008.

[8] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Trans. Speech and Audio Processing*, vol. 3, no. I, pp. 72-82, 1995.

[9] C.M. Bishop, *Pattern Recognition and Machine learning*, Springer Science, New York, 2006.

[10] EMO-DB, "Berlin database of emotional speech," `http://pascal.kgw.tu-berlin.de/emodb/start.html`.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, and W Sendlmeier, "A database of german emotional speech," in *Proc. IEEE Interspeech 2005*, Lisbon, Portugal, 2005.