# SPEAKER RECOGNITION IN AN EMOTIONAL ENVIRONMENT

Marius Vasile Ghiurcau[1], Corneliu Rusu[1] and Jaakko Astola[2]

[1]Signal Processing Group, Faculty of Electronics, Telecommunications and Information Technology
Technical University of Cluj-Napoca
[2]Tampere International Center for Signal Processing
Tampere University of Technology

## ABSTRACT

*The goal of this paper is to assess the effect of emotional state of a speaker when text-independent speaker identification is performed. Mel-frequency cepstral coefficients are the features of the speech signal used for speaker recognition. For training the speaker models and testing the system, Support Vector Machines are employed. Berlin emotional speech database, which contains 10 different speakers recorded in different emotional situations (happy, angry, fear, bored, sad and neutral) is used. The results show an important influence of the emotional state upon text-independent speaker identification. A possible solution to this issue is finally suggested.*

**Keywords:** *speaker identification, emotions, SVM.*

## 1. INTRODUCTION

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. Human beings are not machines and quite frequently are being overwhelmed by emotions such as happiness, fear, anger, sadness, boredom etc. These emotions are present in our everyday life. The studies suggest that approximately 10% of the human life is unemotional [1] while the rest is affected by various emotions; most of the times, one can not really control these emotional states.

Building access systems are very popular lately. Usually these systems are trained using the voice of a certain person in a neutral or normal state. If someone tries to use this kind of access system after a bad day, then his/her voice is rather bored or sad, than neutral. If he/she is being followed on the way back home, its voice is rather attained by fear, scared, not really neutral. Will the identification system still identify and let the user inside? Quite similar situations can be found in different forensic cases, when the voices are being recorded by the surveillance systems. The quality of the recording can be quite good, but even so, may one use this sample in order to make comparisons? Voice may be rather scared, or sad, but not neutral.

Even though one may say that the first two scenarios, concerning the building access systems, could be better assimilated to a speaker verification task, rather than a speaker identification task, we can say that the two tasks are quite similar anyway. Moreover, in terms of signal processing aspects, these problems can be treated almost identically [2].

It should be mentioned that recognition of human emotional state is a topic under development in the last years. So far, most of the work conducted in the *emotional field* implied emotion recognition. Giannakopoulous in [3] suggests a method for extracting affective information using speech data from movies. In [4] it is presented a regression approach to music emotion recognition, with applications in music retrieval field.

As to our knowledge very few work has been done for speaker recognition in emotional environment. Shahin in [5] 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 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. In a recent study [6] we have evaluated the performances of such a such a speaker identification system and the results were not satisfying. Previously the Gaussian Mixture Models (GMMs) were used in the classification process, while in this approach the Support Vector Machines (SVMs) are employed and comparative results shall be presented. Indeed, the present study proposes 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. In the end, we propose a possible solution for increasing the performances of such a speaker identification system and also suggest some other future work possibilities.

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. Section 3 presents the practical work, followed by the results of our research and a short discussion. Some conclusions and future work possibilities can be found in Section 4.

## 2. THEORETICAL BACKGROUND

### 2.1 Mel-frequency cepstral coefficients

The Mel-frequency cepstral coefficients (MFCCs) are probably the most popular features used in sound classification applications. They are a short-term spectrum-based feature which give good discriminative performance.

The extraction of the MFCCs includes some steps that

are going to be briefly presented in the next paragraphs. Firstly, in order to reduce the noise and also to enhance the high-frequency spectrum, a finite-order impulse response (FIR) filter is applied to the audio signal. This filter is called pre-emphasis filter and has the following equation:

$$H_{\text{FIR}}(z) = 1 - az^{-1} \tag{1}$$

The value for $a$ is usually selected from the [0.95, 0.98] interval. After the pre-emphasis, the signal is divided into frames. This framing comes from the necessity of transforming the signal into statistically stationary blocks. Overlapping frames with a 30-50% overlap are used, in order to avoid losing information at the end of the frames. Next comes the windowing process. This is applied in order to prevent abrupt changes at the end points of the frames. Usually a Hamming window is used. For each frame the Discrete Fourier Transform (DFT) is applied. Because humans do not perceive pitch linearly, the frequency band has to be divided using a filter-bank of triangular filters spaced on the Mel-scale. In the end, spectral envelope in dB is obtained by applying logarithm to the amplitude spectrum. Finally, the Discrete Cosine Transform (DCT) is applied [7]:

$$c_n = \sum_{j=1}^{M} \log S_j \cos\left[\frac{\pi n}{M}(j-\frac{1}{2})\right], \quad n = 1,2,...,N, \tag{2}$$

where $c_n$ is the $n^{th}$ MFCC coefficient, $M$ is the number of filterbanks, $N$ is the number of coefficients one wants to compute and $S_j$ is the magnitude response of the $j^{th}$ filterbank channel.

The zeroth coefficient is usually dropped because it is the average log-energy of the frames. Most of the times first and second order differences of the MFCCs are included as a feature. Those are called DELTA and DELTA-DELTA coefficients.

### 2.2 Support Vector Machines

Support Vector Machines represent machine learning algorithms that were first introduced by Vapnick [8]. SVMs were originally proposed for solving binary classification problems, where the main goal is to find an optimal separating hyperplane between the data sets, by maximizing the margins between these classes. The data that is to be classified is transformed into a higher dimensional space using different kernel functions and then the separating hyperplane is constructed.

The problem of binary SVMs can be stated as follows. Given the training samples $(x_i, d_i)$, $i = 1,\ldots,N$ where $x_i$ is an $n$-dimensional input feature vector and $d_i \in \{-1,+1\}$ represent the corresponding labels for the two classes of sounds, the decision function is given by [9]:

$$f(\mathbf{x},\mathbf{w},b) = \text{sign}(\mathbf{w}^T\mathbf{x}+b) \tag{3}$$

where $w$ is an n-dimensional weight vector and b is a scalar multiplier. If the two classes can be linearly separated, the optimal hyperplane can be determined from [9]:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \quad d_i(w_i x_i + b) \geq 1,\ i = 1,\ldots,N. \end{aligned} \tag{4}$$

If the classes cannot be linearly separated, the optimization problem can be reformulated as [9]:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to} \quad d_i(w_i x_i + b) \geq 1-\xi_i,\ \xi_i \geq 0\ i = 1,\ldots,N. \end{aligned} \tag{5}$$

In the above equation $C$ is known as the regularization parameter. Finally, the decision function is given by [9]:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N}\alpha_i d_i k(\mathbf{x},x_i) + b\right) \tag{6}$$

where $\alpha_i$ are called Lagrange multipliers and $k(\mathbf{x},x_i)$ are the Kernel functions.

There are two main options for extending the binary SVM classifier presented to a to multiclass classifier: either by combining more binary classifiers or by taking all the classes at once and directly considering all data in one optimization formulation [10].

## 3. PRACTICAL WORK

### 3.1 Database description

The speech database used in this study is the Berlin Database of Emotional Speech [11] 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) [12]. 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 [12].

### 3.2 Experimental setup

Mel-frequency cepstral coefficients are firstly extracted from the signals. When using SVMs, for training each class of sounds, one needs feature vectors with the same number of components. The recordings from the databases don't have the same length and consequently they have different number of MFCC coefficients, because these coefficients were extracted per frames. In order to solve this issue, the mean value for each MFCC coefficient was computed. Then the training of the models and testing using the Support Vector Machines was employed. For each of the speakers there were available around 10 utterances of a few seconds of speech (in each emotional state). The recordings were sampled at 16 kHz (16 bits mono *.wav files). The pre-emphasis FIR filter had the pre-emphasing coefficient $a = 0.97$. 256-sample frames with an overlap of 128 samples (50% overlap)

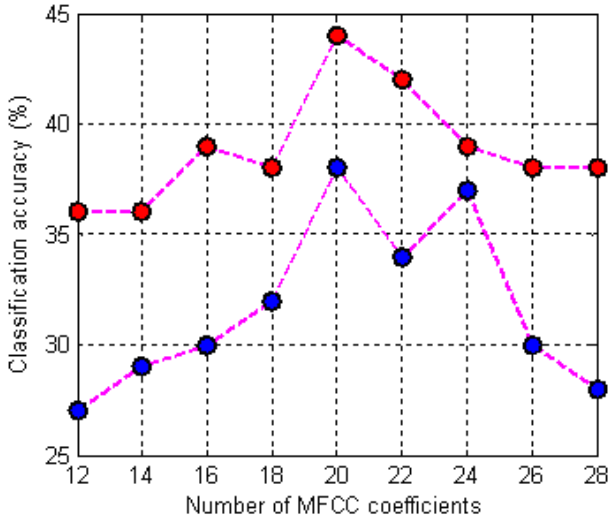


Figure 1*: Results for text-independent speaker identification when training is done with neutral utterances and various emotional states are used for testing. SVM type:* C-SVC *(red) and* nu-SVC *(blue).*

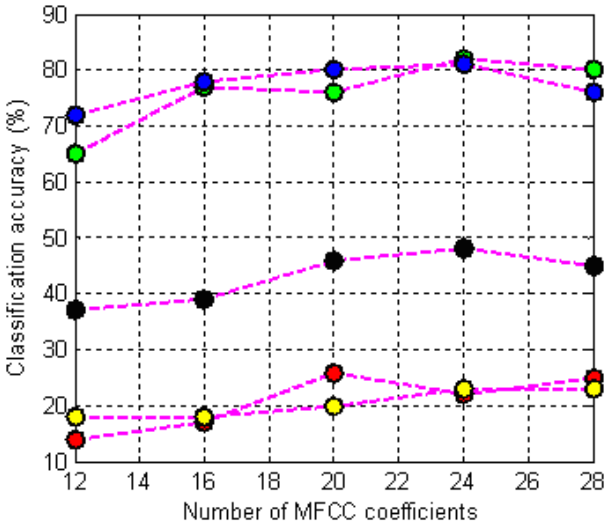


Figure 2*: Results for text-independent speaker identification when training is done with neutral utterances and testing is done using in each case a different emotional state:* sadness *(blue),* boredom *(green),* fear *(black),* anger *(red),* happiness *(yellow)*

were used and the MFCCs were computed for each frame. In these experiments we have tried using 12 to 28 cepstral coefficients and the results were compared. First coefficient was all the time discarded. Along with MFCCs we have also used the DELTA coefficients (first order differences of the MFCCs).

For SVM classification the LIBSVM software [13] was employed. Two types of SVMs were tried, namely C-SVC and nu-SVC. For the Kernel, we have used four different types: *linear*, *polynomial*, *radial basis*, *sigmoid*. The best results were obtained for the polynomial type and only those are going to be presented here.

### 3.3 Results

*Experiment 1*: Ten different utterances of the same emotional state of the speakers were taken. 9 of the utterances were used for training and 1 for testing. This procedure was repeated 10 times, every time choosing another test utterance (leave one out cross validation method). At the end, we computed the average of the results. Overall, depending on the particular emotional state that was simulated, correct identification rates between 99% and 100% were achieved.

*Experiment 2*: The previous experiment was repeated, but this time we have used neutral utterances for the training phase and recordings with different emotional states for the testing phase (for each speaker, we had ten test recordings, 2 recordings of each remaining emotional state). Figure 1 presents the overall correct identification rates for different combinations of the number of MFCCs.

*Experiment 3*: To this end, we trained again the system with neutral utterances and tested it using each time a different emotional state. The results are presented in Figure 2.

*Experiment 4*: 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 SVM model. We tested the system using the remaining recording. This process was repeated 35 times, each time changing the test sample. The overall results are presented in Figure 3.

Increasing the number of MFCCs or changing the corresponding frame length, didn't bring any improvement in the final results. We also tried using only the MFCCs, without the delta coefficients, but the results are not satisfactory.

### 3.4 Comments

When the same emotional state is used for both training and testing the classification scores are very good, with values between 99% and 100%. Those results are similar to the ones obtained previously [6]. Achieving a maximum correct identification rate it is quite normal taking into account the small number of speakers and the fact that closed set identification is performed.

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/SVM combinations managed to get scores above 45% which could be assimilated to a total fail in real applications. Previously the results were better, reaching to scores up to 60%.

As expected, the results of the speaker identification process are strongly dependent on the emotional state. The best results are obtained for boredom and sadness, with scores up to 80% (previously: 89%). The worst scores are obtained for both anger and happiness, none of them managing to get over 26% (previously: 36%). For fear, the best result was 48% (previously slightly better: 52%).

Finally, it seems that if we train the system with all the available emotional states and also test it with different emo-

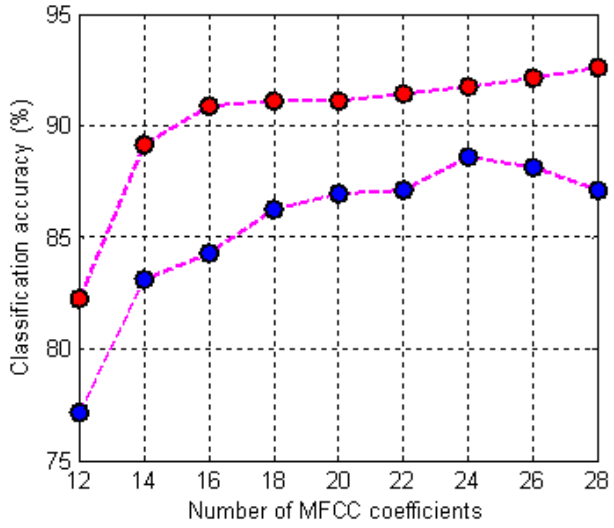


Figure 3*: Text-independent speaker identification when both training and testing is performed using various emotional states. SVM type:* C-SVC *(red) and* nu-SVC *(blue).*

tional states the results increase significantly, to scores of around 92%. The best score of 92.57% was achieved for C-SVC and 28 MFCC coefficients. When the GMM approach was used, the best rates were higher (98.57%). It seems that, in all the experiments, the results obtained with GMM are better than the ones obtained with SVMs. Even so, neither of them can be considered satisfactory.

## 4. CONCLUSIONS AND FUTURE WORK

As it was expected, MFCC and SVM perform well in text-independent speaker identification. When emotions alter the human voice, the performances of the speaker recognition systems decrease significantly. If one trains the systems with all the emotional states available, the correct classification rates increase, with values that reach up to approximately 93%. 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%.

We consider that training the system with utterances in different emotional state can be just a theoretical solution; creating a database for regular (not actors) users in different emotional states is not really a viable solution, due to the difficulties encountered by users when asked to simulate emotions. Finding other speech features that are not strongly dependent on the emotional state, or a combination of both GMM and SVM could be tried in the future.

## REFERENCES

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

[2] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[3] 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.

[4] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. 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.

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

[6] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *Proc. ICASSP 2011*, Prague, Czech Republic, 2011.

[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The Book (for HTK V3.1)*. Cambridge University, 2000.

[8] V. Vapnick, *The Nature of Statistical Learning Theory*. Springer-Verlag, New-York, 1995.

[9] S. Chandaka, A. Chatterjee, and S. Munshi, "Support vector machines employing cross-correlation for emotional speech recognition," *Measurement*, vol. 42, no. 4, pp. 611 – 618, 2009.

[10] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," in *IEEE Transactions on Neural Networks*, vol. 3. no. 2, pp. 415-425, Mar. 2002.

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

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

[13] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

**Acknowledgements**

The work of Marius Ghiurcau was supported by PRODOC POSDRU/6/1.5/S/5 ID 7676. The work of Corneliu Rusu was supported by CNCSIS Grant ID 162/2008.