Journal of Engineering Science and Technology
Vol. 11, No. 9 (2016) 1221 - 1233

COMPARISON BETWEEN GMM-SVM SEQUENCE KERNEL AND GMM: APPLICATION TO SPEECH EMOTION RECOGNITION

I. TRABELSI*, D. BEN AYED, N. ELLOUZE

Université Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis-ENIT
Laboratoire Signal, Image et Technologies de l'Information-LRSITI, 1002, Tunis, Tunisia
*Corresponding Author: imen.trabelsi@enit.rnu.tn

Abstract

Speech emotion recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional state is an important factor in human communication, because it provides feedback information in many applications. This paper makes a comparison of two standard methods used for speaker recognition and verification: Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) for emotion recognition. An extensive comparison of two methods: GMM and GMM/SVM sequence kernel is conducted. The main goal here is to analyze and compare influence of initial setting of parameters such as number of mixture components, used number of iterations and volume of training data for these two methods. Experimental studies are performed over the Berlin Emotional Database, expressing different emotions, in German language. The emotions used in this study are anger, fear, joy, boredom, neutral, disgust, and sadness. Experimental results show the effectiveness of the combination of GMM and SVM in order to classify sound data sequences when compared to systems based on GMM.

Keywords: Speech, Emotions, SVM, GMM, Kernel, Sequence.

1. Introduction

Recognition of emotions in speech is essential for understanding human interactions and hence is a complex task that is furthermore complicated because there is no unambiguous answer to what the "correct" emotion is for a given speech sample. Although numerous feature extraction schemes based on both acoustic and prosodic features have been used [1-4] and different pattern recognition methods as Artificial Neural Networks (ANN) [5], Hidden Markov

Nomenclatures

b_i	Component densities
d	Learned constant
Pi	Mixture weights
xi	Support vectors

Greek Symbols

α_i	Lagrange multiplier.
Φ	Kernel mapping
γ	Free shape parameter
τ	Relevance factor

Abbreviations

ANN	Artificial Neural Networks
EM	Expectation Maximization
GMM	Gaussian Mixture Model
RBF	Radial basis function
SVM	Support Vector Machines
UBM	Universal Background Model

Model (HMM) [5, 6], Gaussian Mixture Model (GMM) [7] have been developed, the emotion recognition accuracy is still short of what is desired. This paper describes the combination of both methods GMM which are a generative model and the discrimination power of SVM via the use of sequence discriminant kernels. Analysis of the behaviour of such hybrid systems is the object of this study. The evaluation presented in this work has the following parameters: (1) 7 emotions from the Berlin database of emotional speech (Emo-Db), (2) Text-independency and speaker-independency were assumed, and (3) Feature vectors were extracted on frame level.

The rest of this paper is organized as follows. The GMM standard approach is presented in Section 2. In Section 3, there is an outline of the used combination of GMM and SVM. In Section 4, the used emotional database, the employed speech features and the test protocol are presented. Section 5 presents and discusses the research results obtained. Conclusions are drawn in section 6.

2. Basic Principles of Applied Classification Method GMM Emotion Discrete Classification

Gaussian mixture density techniques are used to learn the extracted features from speech. Our assumption is that we have enough data to robustly estimate weight, mean and variance parameters for each emotion class individually. In the recognition phase, testing unknown emotional samples are used to evaluate the performances of models. Posterior probability of the features of a given speech utterance is maximized over all emotion GMM densities. The GMM emotional model that has the maximum log-likelihood with a given input utterance is determined to be the recognized emotional state. Figure 1 shows the speech emotion classification scheme based on GMM.

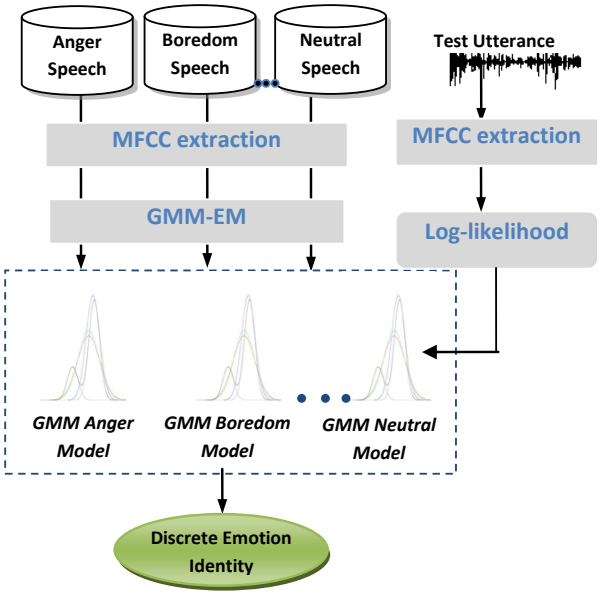


Fig. 1. GMM emotion discrete classification.

Gaussian mixture model is a probabilistic model for density estimation using a convex combination of multi-variate normal densities. A GMM aims to approximate a complex nonlinear distribution using a mixture of simple Gaussian models, each parameterized by its mean vector, covariance matrix and the mixing parameter. These parameters are learned by the iterative Expectation Maximization technique EM [8] and initialized by a clustering algorithm. The EM algorithm iteratively renews the GMM parameters in order to increase the likelihood of the estimated model for the observed GMM parameters. This approach alternates between performing an Expectation step (E-step), which computes the distribution for the hidden variables using the current estimates for the parameters, and a Maximization step (M-step), which re-estimates parameters to be those maximizing the likelihood found in the E-step.

The Gaussian mixture density is given by:

$$p(x|\lambda) = \sum_{i=1}^{N} P_i \, b_i(x) \tag{1}$$

where x is a dimensional random vector, $b_i(x)$ are the component densities and pi the mixture weights. Each component density is a d-variate Gaussian function having the form:

$$b_i(x) = \frac{1}{(2\Pi)^{\frac{d}{2}} \left| \sum_i \right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i) \right] \tag{2}$$

Mixture weights (p_i) must satisfy constraint:

$$\sum_{i=1}^{N} p_i = 1 \tag{3}$$

The number of Gaussian components indicates the number of clusters within each class. The GMM classifier returns probabilities that the tested utterance belongs to the GMM model, which is trained for each emotion category. The best matching emotion is given by the maximum overall probability for the given unknown emotion. Two critical factors in training a Gaussian mixture emotional model are selecting the order of the mixture and initializing the model parameters prior to the EM algorithm.

3. Basic Principles of Applied Classification Method GMM-SVM Discriminant Sequence Kernels

A hybrid classifier based on the combination of a generative model GMM and discriminative classifier SVM is proposed, to achieve better classification and computation performances. The GMM-SVM paradigm has been introduced and successfully applied to speaker recognition [9]. In this work, GMM supervector based SVM is adopted for emotion recognition of speech. Details of the method are described as follows.

3.1. Support vector machines and discriminant sequence kernels

The support vector machine (SVM) [7] performs essentially a binary nonlinear classification based on hyperplane separation. SVM performs a non-linear mapping from an input space to a high-dimensional space thanks to kernel functions to achieve a maximum margin hyperplane.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + d \tag{4}$$

where x_i are the support vectors chosen from training data via an optimization process and $yi \in \{-1, +1\}$ are respectively their associated labels. N denotes the number of support vectors and d is a (learned) constant. $K(x, x_i)$ is the kernel function and must fulfill some conditions:

$$K(x, y) = \Phi(x)^t \Phi(y) \tag{5}$$

where Φ is a mapping from the input space to a possible infinite-dimensional space.

The use of SVM at frame level, as in the case of GMMs, showed its limitations for speech recognition [10] in terms of efficiency and training time, as the number of frames increases. This can be overcome by using SVM kernels to classify sequences instead of frames. The used kernel in this study is the radial basis function (RBF) kernel. LIBSVM package [11] is used for SVM training and testing.

$$K(x_i, x_j) = \exp\left[-\gamma \left| x_i - x_j \right|^2 \right] \tag{6}$$

3.2. GMM supervector

First, a universal background model (UBM) is learned with multiple audio files from all training different emotions. The UBM is trained with the EM algorithm on its training data. From this initial model, an adapted GMM is created for each

emotional class by maximum a posteriori (MAP) estimation, allowing for prior distribution to be incorporated in the final estimation process. This allows a detailed model to be trained when little data is available, which is often the case when a large number of parameters are estimated. Only the mean vectors are adapted while the covariance matrices and weights remain unchanged. The prior distribution for this estimation is determined by the UBM parameter and a factor τ governing the influence or relevance of the UBM on the final emotion model. The relevance factor is a way of controlling how much observed training data influence the model adaptation. From the adapted GMM, the final GMM supervector is constructed as the representation of the input utterance. The supervector of a GMM is defined by concatenating the mean of each Gaussian mixture, which can be thought of as a mapping between an utterance and a high-dimensional vector. The framework of the proposed emotion recognition system is illustrated in Fig. 2.

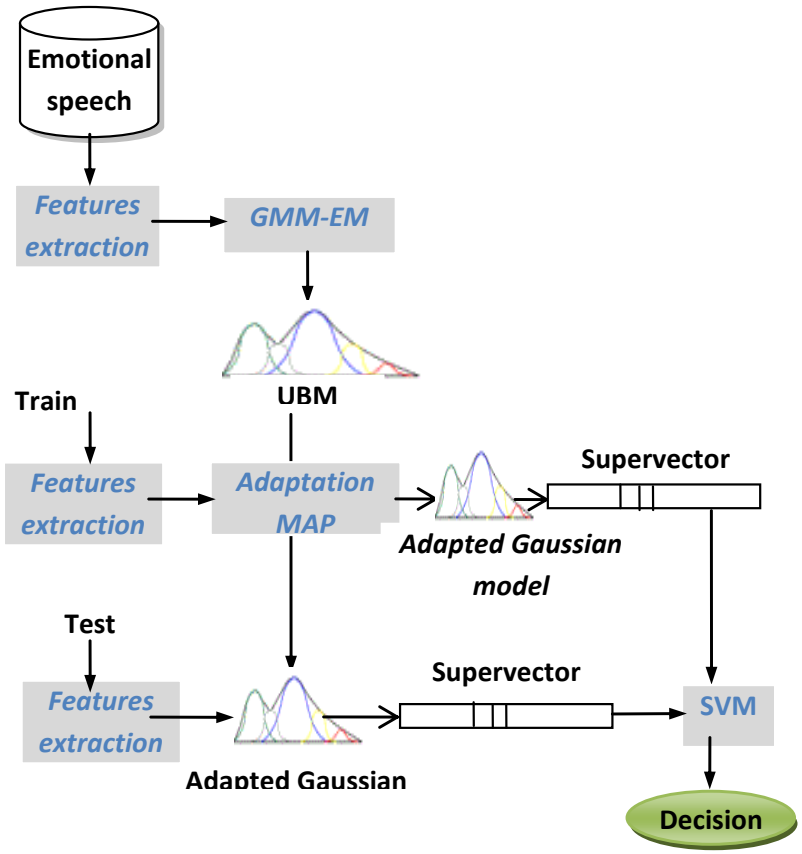


Fig. 2. GMM-SVM emotion discrete classification.

4. Experimental Evaluation

4.1. Database

In EMO-DB [12], ten professional native German actors (5 female and 5 male) simulated 7 emotions, producing 10 utterances. 5 utterances are short, while the

remaining 5 are long. The emotions are: anger, boredom, disgust, fear, happiness, sadness, and neutral [12]. This emotional speech corpus is probably the most often used database in the context of emotion recognition from speech, and also one of the few for which some results can be compared. The speech data were recorded at a sample rate of 16 kHz and a resolution of 16 bits. In this study, only 500 sentences are used. These sentences were not equally distributed between the various emotional states as shown in Table 1.

Table 1. Number of utterances in Emo-Db.

Emotion	Number
Anger	128
Boredom	81
Disgust	44
Fear	69
Joy	71
Sadness	62
Neutral	45

The whole available dataset has been divided into two subsets: one with 70 percent of the source data, for training the model, and one with 30 percent of the source data, for testing the model.

4.2. System description

The classification performance largely relies on the kind of features we can extract. Features such as mel frequency cepstral coefficients (MFCCs) are typically used in automatic speech recognition, as these cater for a robust and reliable recognition performance independent of the speakers and the accompanying different characteristics of voices which also change according of the speakers' emotional states. In [13, 14], the authors showed that MFCC features yield surprisingly good performance on the emotion recognition task. In this paper, MFCC are adopted as feature parameters for our SER. First, the signal is passed through a pre-processing system that normalizes amplitude, reduces the amount of noise and extracts MFCC parameters. Mel filter banks are placed in [20-3000] Hz. The used evaluation measure is the unweighted accuracy (UA), i. e., the unweighted average of the recalls of the 'positive' and 'negative' classes, which has been the official competition measure of the first of its kind INTERSPEECH 2009 Emotion Challenge [15].

5. Results and Discussion

5.1. GMM emotion discrete classification results

The experiments below were aimed at comparison and analysis of:

- ✓ Influence of different training size on GMM emotion classification;
- ✓ Influence of the used GMM components number;

- ✓ Influence of the used EM iterations on the GMM emotion classification;
- ✓ Influence of the matrix covariance choice;
- ✓ Influence of the used initialization algorithm;
- ✓ Influence on the sorted GMM mixtures on the GM emotion classification.

Experiment1- Relation between emotion recognition performance and different amount of training data

Figure 3 shows the effect of the data set size on the classification performance for the seven emotions. As it can be seen in the figure, larger amounts of training data help to increase emotion recognition rate. But also, when the training set is sufficiently large, additional training data makes less impact on increasing the emotion recognition rate.

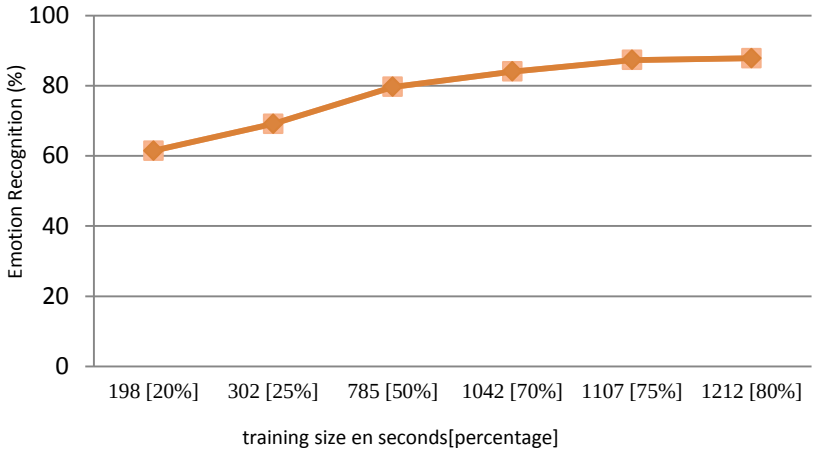


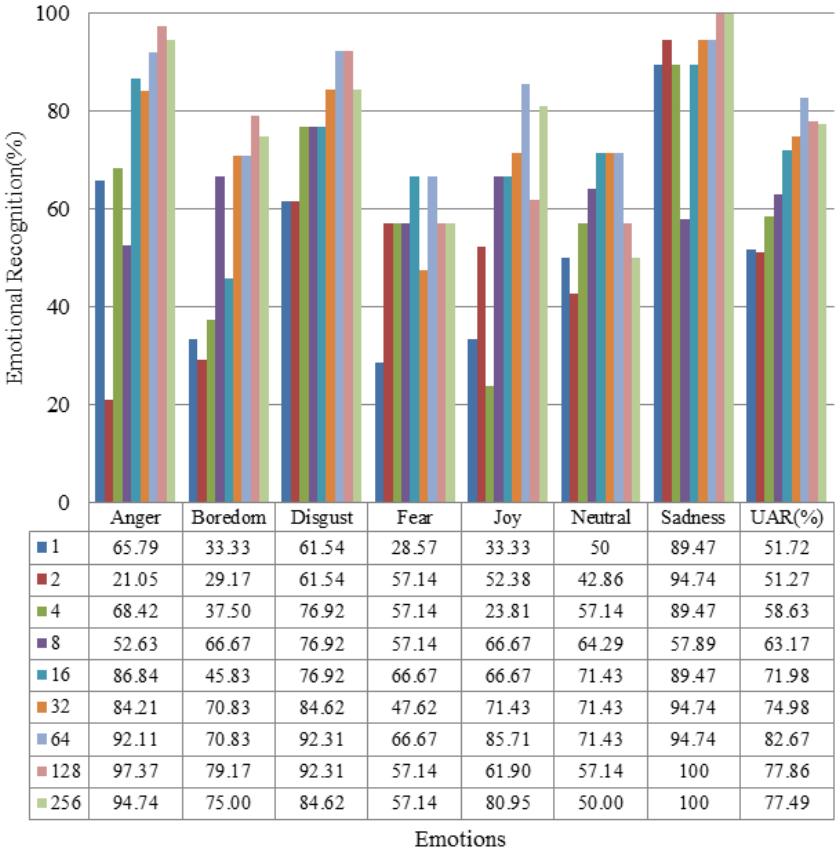
Fig. 3. Relationship between emotion recognition rates and various training sizes.

Experiment 2 - Choosing the number of GMM components.

There is no theoretical way to determine the number of mixture components (model order) and the optimum number of training iterations (EM iterations) a priori. The selection of the model order is a compromise between the amount of training data and the resolution of GMM modeling ability. This experiment has to find out the best number of mixtures which corresponds to the best system performance. *N* mixtures are conveniently picked, going from 1 to 256 with diagonal covariance matrices and one EM iteration for all GMM based density functions. Change of the recognition rate along with the number of Gaussian mixture components N is observed and shown in Fig. 4.

The obtained results showed that too few components will not be able to accurately model the distinguished characteristics of an emotion distribution. Another thing to note is that from 1 mixture up to 64 mixtures, the accuracy increases in small steps, but from 64 to 256 mixtures a small drop is seen. We

remark also that sadness is well recognized even with one GMM mixture. Too many components relative to limited training data induce too many free parameters to be estimated reliably, thus degrade performance. For this task, the highest recognition rate of 82. 7 % is achieved by 64 GMM mixtures.



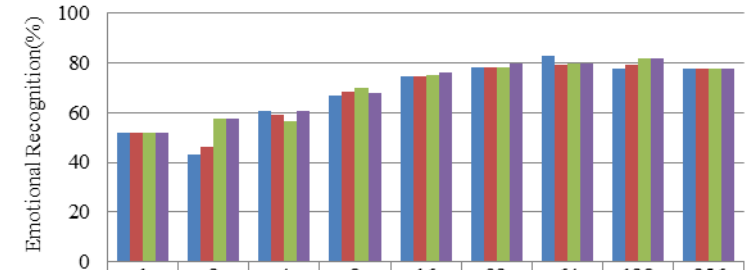
	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	UAR(%)
1	65.79	33.33	61.54	28.57	33.33	50	89.47	51.72
2	21.05	29.17	61.54	57.14	52.38	42.86	94.74	51.27
4	68.42	37.50	76.92	57.14	23.81	57.14	89.47	58.63
8	52.63	66.67	76.92	57.14	66.67	64.29	57.89	63.17
16	86.84	45.83	76.92	66.67	66.67	71.43	89.47	71.98
32	84.21	70.83	84.62	47.62	71.43	71.43	94.74	74.98
64	92.11	70.83	92.31	66.67	85.71	71.43	94.74	82.67
128	97.37	79.17	92.31	57.14	61.90	57.14	100	77.86
256	94.74	75.00	84.62	57.14	80.95	50.00	100	77.49

Fig. 4. Influence of the used number of mixture on GMM recognition.

Experiment 3 - trials with number of iterations using EM algorithm

An exhaustive series of preliminary results varying the number of mixtures and EM iterations are processed. A selection of these results is provided in Fig. 5.

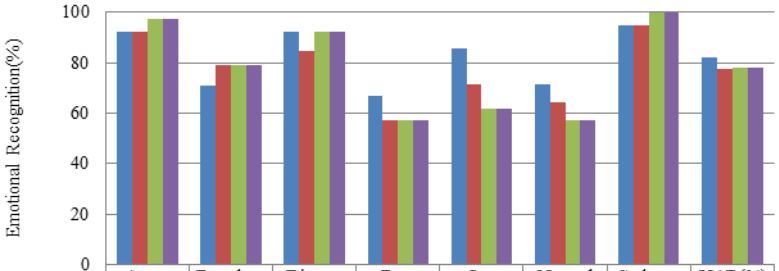
No significant difference of emotion recognition rate was found between 1 and 1000 iterations for all GMM components. When one GMM mixture is utilized, varying the EM iterations has no impact on the accuracy of the system. The same thing is noted when 256 mixtures are employed. The highest classification accuracy was obtained with 64 mixtures and one EM mixture. Figure 6 shows the comparison of classification accuracies of all emotions, in the case of 64 mixtures.



	1	2	4	8	16	32	64	128	256
1EM	52	43.33	60.67	66.67	74.67	78.00	82.70	77.86	77.49
10EM	52	46	59.33	68.66	74.66	78.00	79.33	79.33	77.49
100EM	52	57.33	56.66	70	75.33	78	80	82	77.49
1000EM	52	57.33	60.67	68	76	80	80	82	77.49

EM iterations

Fig. 5. Influence of the used number of EM iterations on GMM recognition.



	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	UAR(%)
1EM	92.11	70.83	92.31	66.67	85.71	71.43	94.74	81.97
10EM	92.11	79.17	84.62	57.14	71.43	64.29	94.74	77.64
100EM	97.37	79.17	92.31	57.14	61.90	57.14	100	77.86
1000EM	97.37	79.17	92.31	57.14	61.90	57.14	100	77.86

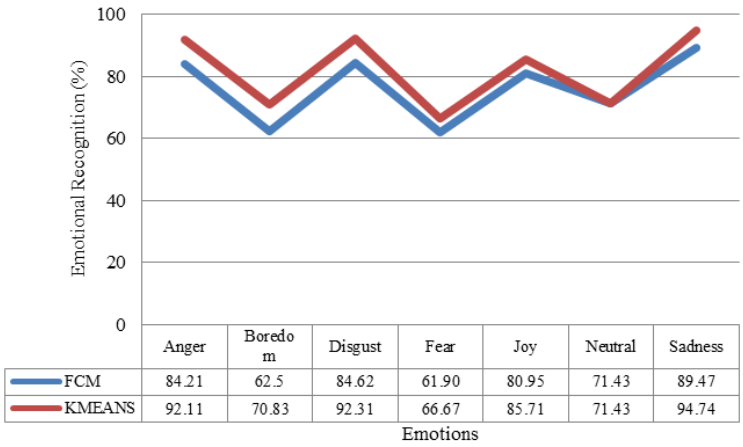
Emotions

Fig. 6. Influence of the used number of EM iterations on all emotions.

Experiment 4 -Influence of the used initialization algorithm: K-means vs Fuzzy-C-Means

The parameters of the GMM for emotional models, in general, are estimated iteratively using the EM algorithm, which converges to the maximum likelihood estimate of the mixture parameters. However, the EM optimization strategy has been known to suffer from several problems. One of the problems is that, as a local method, it is too sensitive to the initial set-ups of the parameters and it may converge to the boundary of parameter leading to inaccurate estimation.

To verify this problem, Fig. 7 shows the influence of the initialization algorithm and compares the performance of a soft clustering (e.g., fuzzy C means or FCM) and a hard clustering technique (e.g., K-means or KM). The obtained results show that K-means algorithm is better than FCM algorithm.



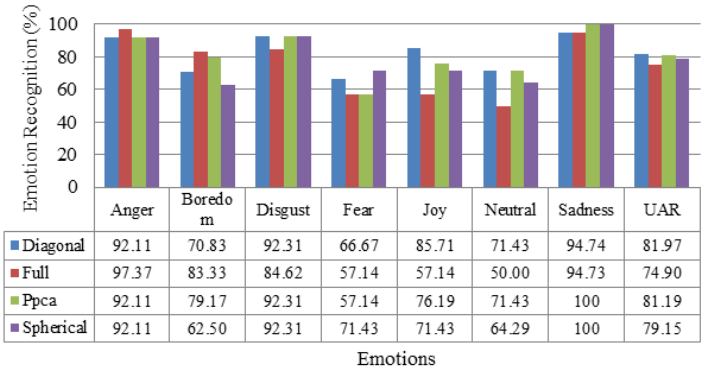
	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
FCM	84.21	62.5	84.62	61.90	80.95	71.43	89.47
KMEANS	92.11	70.83	92.31	66.67	85.71	71.43	94.74

Fig. 7. Comparative analysis of K-means and FCM.

Experiment 5 - Influence of the matrix covariance choice

Commonly, GMM models in speech recognition applications assume diagonal covariance matrices. This experiment takes different covariance matrix approximations. We compared the performance of diagonal, full, spherical and probabilistic principal component analyzers (PPCA) covariance matrices. The results are shown in Fig. 8.

This experiment reveals that in this task, the diagonal covariance models perform better than the other covariance models.



	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	UAR
Diagonal	92.11	70.83	92.31	66.67	85.71	71.43	94.74	81.97
Full	97.37	83.33	84.62	57.14	57.14	50.00	94.73	74.90
Ppca	92.11	79.17	92.31	57.14	76.19	71.43	100	81.19
Spherical	92.11	62.50	92.31	71.43	71.43	64.29	100	79.15

Fig. 8. Comparative analysis of matrices covariance.

5.2. Method GMM-SVM discriminant sequences Gaussian kernel results

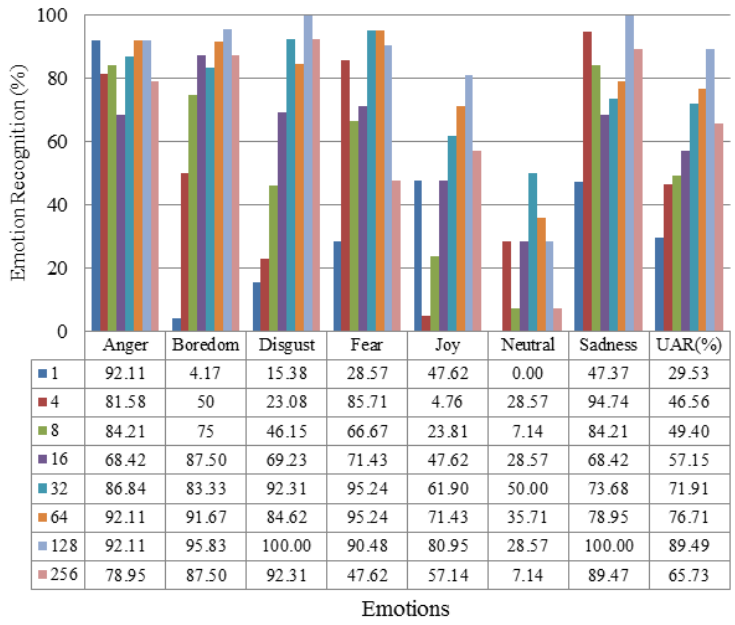
The experiments conducted in this section were aimed at comparison and analysis of:

- ✓ Influence of the used GMM components number;
- ✓ Influence of the used relevance factor number;
- ✓ Influence of the used EM iterations on the GMM emotion classification;

- ✓ Influence on the sorted GMM mixtures on the GM emotion classification.

Experiment 1 - trials with different number of Gaussian components

Figure 9 reports results for different Gaussian component numbers and found out that the optimal dimensions for this system is 128.



	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	UAR(%)
1	92.11	4.17	15.38	28.57	47.62	0.00	47.37	29.53
4	81.58	50	23.08	85.71	4.76	28.57	94.74	46.56
8	84.21	75	46.15	66.67	23.81	7.14	84.21	49.40
16	68.42	87.50	69.23	71.43	47.62	28.57	68.42	57.15
32	86.84	83.33	92.31	95.24	61.90	50.00	73.68	71.91
64	92.11	91.67	84.62	95.24	71.43	35.71	78.95	76.71
128	92.11	95.83	100.00	90.48	80.95	28.57	100.00	89.49
256	78.95	87.50	92.31	47.62	57.14	7.14	89.47	65.73

Fig. 9. Influence of the used number of mixture on GMM recognition.

Experiment 2 - trial with different values of relevance factor

The GMM-SVM classifier is evaluated with seven different relevance factor values. From the results presented in Fig. 10, the optimal relevance factor value is 16.

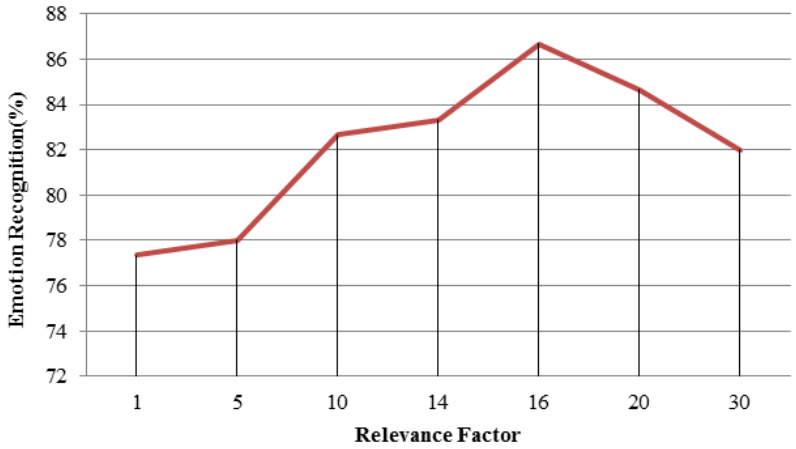


Fig. 10. Influence of the used relevance factor.

5. Conclusions

Automatic recognition of emotions used in a text independent speech and speaker independent environment has been evaluated using MFCC features and were conducted on the actor recorded simulated database Emo-DB. The modeling step in speech emotion recognition has an enormous influence on the recognition task. In fact, the initial model parameters have an influence on the final determined parameters of the emotion models. The experiments were aimed at comparison and analysis of several parameters in two methods. The first one provided a performance evaluation of the GMM model. The objective of this study is to define the optimal GMM parameters. Such parameters were the optimum number of Gaussian densities, optimum number of EM iterations, and the covariance matrix type. We tried also to see the impact of selecting GMM with the N top highest weights on the emotion recognition performance. The second one is based on a hybrid GMM/SVM approach. Significant improvement in speech emotion recognition is achieved. It can be easily observed that GMM/ SVM has provided a better performance than GMM for emotion identification task.

Although several papers have been published on emotional speech recognition, very few have considered hybrid classification methods. In [16], a hybrid classification method that combined the decision trees and the GMM means supervector was proposed. The highest accuracy classifying three emotional classes was reported to be 84%. In [17], the Linear Discriminant Classifier with Gaussian class-conditional probability distribution and K-Nearest Neighbors were used to recognize negative and non-negative emotional classes. The highest accuracy was reported to be 89.06%. In [18], the authors proposed a hybrid scheme that combines the Probabilistic Neural Network (PNN) and the GMM for identifying emotions from speech signals. Experimental results showed that the proposed scheme were able to achieve 80.75%. The method explored in this paper achieved a recognition rate of 89.49% to recognize seven different emotional classes.

Considering the fact that the current used database consists of speech only with acted emotional styles, we plan to develop an automatic emotion recognizer in real situations.

References

1. El Ayadi, M.; Kamel. M.S.; and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44 (3), 572-587.
2. Lugger, M.; and Yang, B. (2008). Cascaded emotion classification via psychological emotion dimensions using large set of voice quality parameters. *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Las Vegas, Nevada, U.S.A, 4945–4948.
3. Rong, J.; Li, G.; and Phoebe Chen, Y.-P. (2009). Acoustic feature selection for automatic emotion recognition from speech. Information Processing and Management, 45(3), 315–28.

4. Zeng, Z.; Pantic, M.; Roisman, G.I.; and Huang, T.S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39-58.
5. Mao, X.; Chen, L.; and Fu, L. (2009). Multi-level speech emotion recognition based on HMM and ANN. *Proceedings of 2009 WRI World Congress on Computer Science and Information Engineering*,7, 225-229.
6. Sethu, V.; Ambikairajah, E.; and Epps, J. (2013). On the use of speech parameter contours for emotion recognition. *EURASIP Journal on Audio, Speech and Music Processing*, 13(1), 1-14.
7. Steinwart, I.; and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.
8. Dempster, A.P.; Laid, N.M.; and Durbin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.
9. Trabelsi, I.; and Ben Ayed, D. (2012). On the use of different feature extraction methods for linear and non linear kernels. *Proceedings of the 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, 797-802.
10. Amami, R.; Ben Ayed, D.; and Ellouze N. (2014). Incorporating belief function in SVM for phoneme recognition. *Hybrid Artificial Intelligence Systems, Lecture Notes in Computer Science*, 8480, 191-199.
11. Chang, C.-C.; and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.
12. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; and Weiss, B. (2005). A database of German emotional speech. *Proceedings of Interspeech*, Lisbon, Portugal, 1517-1520.
13. Ser, W.; Cen, L.; and Yu, Z.-L. (2008). A hybrid PNN-GMM classification scheme for speech emotion recognition. *Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008)*, 1-4.
14. Trabelsi, I.; Ben Ayed, D.; and Ellouze, N. (2013). Improved frame level features and SVM supervectors approach for the recogniton of emotional states from speech: Application to categorical and dimensional states. *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, 5(9), 8-13.
15. Trabelsi, I.; and Ben Ayed, D. (2013). A multi level data fusion approach for speaker identification on telephone speech. *International Journal of Signal Processing, Image Processing & Pattern Recognition*, 6(2), 33-42.
16. Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D. (2011). Recognizing realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10), 1062-1087.
17. Garg, V.; Kumar, H.; and Sinha, R. (2013). Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. *Proceedings on 2013 National Conference on Communications (NCC)*, 1-5.
18. Lee, C.-M.; Narayanan, S.; and Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. *ASRU'01. IEEE Workshop on Automatic Speech Recognition and Understanding*, 240-243.