Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients

Farah Chenchah
LR-SITI Laboratory
National Institute of Applied Science and Technology,BP.676 centre urbain cedex
Tunis, Tunisia

Zied Lachiri
LR-SITI Laboratory
National Institute of Applied Science and Technology,BP.676 centre urbain cedex
Tunis, Tunisia

***Abstract*—Recognizing human emotions through vocal channel has gained increased attention recently. In this paper, we study how used features, and classifiers impact recognition accuracy of emotions present in speech. Four emotional states are considered for classification of emotions from speech in this work. For this aim, features are extracted from audio characteristics of emotional speech using Linear Frequency Cepstral Coefficients (LFCC) and Mel-Frequency Cepstral Coefficients (MFCC). Further, these features are classified using Hidden Markov Model (HMM) and Support Vector Machine (SVM).**

Keywords—Mel Frequency Cepstral Coefficients (MFCC); Linear Frequency Cepstral Coefficients (LFCC);Hidden Markov Model (HMM); Support Vector Machine (SVM); emotion recognition

I. INTRODUCTION

Emotion is an important aspect of human interaction that needs to be further investigated. Its understanding becomes essential for understanding human communication. Studies on emotion involve several field of research such psychology, neuroscience, philosophy, physiology, computer science and in several other areas. This melting pot of discipline studying emotion gives to emotion recognition all its importance.

Emotions are a whole-body phenomena that are reflected through several cues such facial expression, body gesture and speech. In this context, advances in automatic speech recognition (ASR) have consumed tremendous effort and have reached a level of maturity which results may be widely used in recognizing emotion.

Speech processing techniques provide an extensive array of feature extraction methods that may be used to extract emotional characteristic in human voice. These features can be divided into two main classes: prosodic and spectral features. Prosodic features include but are not limited to Pitch, Energy, Formant frequency[1], Jitter, Shimmer[2], Zero Crossing Rate (ZCR)[3]. Among spectral features, we can list Linear Predictive Coding (LPC)[4], short-time coherence method (SMC) [5] and Mel-Frequency Cepstral Coefficients (MFCC) [6].

Furthermore, extensive work on emotion recognition has been carried out using different classifier such us neural networks[7], Support Vector Machine (SVM)[8], Gaussian mixture models (GMM) [9] and Hidden Markov Model (HMM) [10].

The aim of this paper is to determine which of Hidden Markov Models (HMM) and Support Vectors Machines (SVM) as classifier and MFFC and LFCC, as feature extraction method can be used to derive an efficient system of emotion through vocal channel. The rest of this paper is organized as follows. In Section 2, system design and selected corpora are presented. Then feature extractions are introduced in section 3and classification models are described in Section 4. Experiments and results are presented in Section 5. Finally, Section 6 gives the conclusions.

II. SYSTEM DESIGN AND SELECTED CORPORA

A. System design

The proposed emotion recognition system can be divided into two main parts, feature extraction and emotion classifier. In the feature extraction, we extract all the acoustical features from both of training and testing speeches. Step classification is performed using Hidden Markov Models (HMM) and Support vector Machines (SVM) to identify the emotion class of a speech utterance. System description is illustrated in figure1.

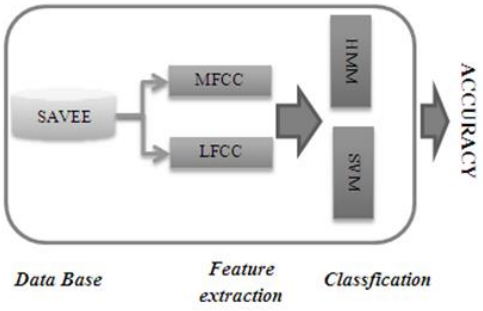

Fig. 1. System description

B. Selected corpus(SAVEE database)

The Surrey Audio-Visual Expressed Emotion (SAVEE) database is a freely available audiovisual data set [11]. This English-language corpus consists of 480 phonetically balanced utterances spoken by four native British male speakers (DC, JE, JK, DC) in seven different emotions (fear, anger, disgust, sadness, surprise, happiness, neutral).

Recordings consisted of 15 TIMIT sentences per emotion (with additional 30 sentences for neutral state).The emotion assessment of recordings was performed by subjective evaluation under audio, visual and audio-visual scenarios. Speech data were labeled at phone-level in a semi-automated way. The sampling rate used for audio data is 44.1 kHz.

III. Feature Extraction

Extracting suitable features from signal is an important step in emotion recognition system. Significant descriptors can carries large emotional information about the speech signal; they affectively increase the performance of classifiers. Several researches have shown that effective parameters to distinguish a particular emotional state with potentially high efficiency are spectral features such as Linear Frequency Cepstral Coefficients (LFCC) and Mel Frequency Cepstral Coefficients (MFFC).

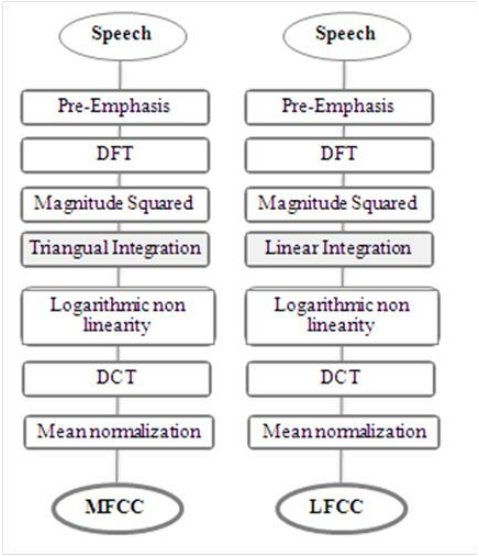

Fig. 2. Mel and Linear filter banks

A. Linear Frequency Cepstral Coefficients (LFCC)

In feature extraction process, [12] introduce a feature method called Linear Frequency Cepstral Coefficients.

The computation of LFCC features can be described; firstly, Fast Fourier Transform (FFT) is applied to windowed signal for converting each frame of N samples from the time domain into the frequency domain. After the FFT block, the power coefficients are filtered by linear frequency filter banks. Finally, the log Mel spectrum is converted into time using Discrete Cosine Transform (DCT).

B. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC), introduced by [13], are based on human hearing perceptions which cannot perceive frequencies over 1Khz.

As presented in figure 2, compute MFCC is similar to LFCC with only difference in the spacing of the filter bank For MFCC, after the FFT block, the power coefficients are filtered by a triangular band pass filter bank also known as Mel-scale filter. The Mel-frequency scale is linear frequency spacing below 1 kHz and logarithmic spacing above 1 kHz.

IV. Classification

In general, there are two approaches to develop classifiers: a parametric approach, and a nonparametric approach. This study uses two nonparametric approaches which are Support vector machines (SVM), often called kernel-based methods and Hidden Markov Model (HMM).

A. Hidden Markov Models (HMM)

A Hidden Markov Model is a doubly stochastic modeling appropriate for no stationary stochastic sequence [14]. HMMs lie at the heart of emotion recognition through vocal channel systems.

HMM is a variant of a finite state machine defined by a (i) set of hidden states, (ii) a transition probabilities distribution, (iii) observation symbol probability distribution in each state and (iv) initial state probability distribution.

The aim of the training phase of the HMM is to decide which one of the H MMs is more likely to have emitted the received sequence. For this purpose, the parameters describing an HMM are estimated. To this end, one or several observation sequence that has been generated by the corresponding stochastic process is used to estimate the unknown parameters.

B. Support Vector Machines (SVM)

Support Vector Machine is a very efficient and simple classifier algorithm which is widely used for pattern recognition.

SVM is a non-linear classifier by transforming the original input set into a higher dimensional feature space by using kernel mapping function, it searches fort the linear optimal separating hyper plane [15].

The most frequently used SVM kernel functions are linear kernel, polynomial kernel and Radial Basis function (RBF) kernel. Considering data from two different classes, an SVM attempts to solve an optimization problem that finds a hyper plane that separates the data with maximum margin. The binary class problem is extended to multiclass classification, methods such as One-Against All (OAA) and One-Against-One (OAO) can be applied.

OAA is the earliest and simplest approach. It involves k binary SVM classifiers, one for each class. Each binary SVM is trained to separate one class from the rest. The winning class is the one that corresponds to the SVM with highest output. OAO involves $k(k-1)/2$ binary SVM classifiers. Each classifier is trained to separate each pair of classes. There are different strategies used to combine these binary classifiers. The crucial widely used strategy is a majority voting.

V. Experimental Setup and Results

A. Experimental setup

In our experimental studies, we collect all the available sentences which are classified in four emotional states that we examine: angry, happy, neutral and sad.

The utterances are expressed by 4 male actors. The sampling frequency of each recording is 16KHz.Signals samples are segmented into frames with 50% overlap.

The feature vector of MFCC and LFCC consists of 13 coefficients. Extraction of cepstral coefficients from emotional speech was performed using LFCC-RASTAMAT toolbox. To compare the effectiveness of these features, step classification was performed using HMM and SVM.

The first step consist of varying the frame length in the range of {50ms, 100ms, 250ms, 500ms, 750ms, 900ms, 1s}. Data was tested using MFCC as feature vector and HMM as classifier. The best frame length obtained will be used for the remaining experiments.

For HMM classifier, we evaluate the topology by varying the number of mixture components and the number of states. HMM models are built for four emotions individually. The HMM classification is done using the Hidden Markov Toolkit (HTK) [16].

For SVM classifier, two Kernel functions are used, polynomial and gaussian, with multiclass strategies, OAA and OAO. To select suitable parameters for each Kernel (C,σ), a cross-validation algorithm was performed by varying the regularization parameter C in [1,100] and Gaussian width σ in [1,10]. The SVM classification is done with the SVM-KM Toolbox for Matlab [17].

B. Results

Speech emotion recognition is implemented using MFCC and LFCC features, we evaluate the system of recognizing emotion state with two classifiers using HMM and SVM.

TABLE I. CLASSIFICATION ACCURACY USING HMM AND MFCC USING DIFFERENT FRAMES

Frame	**0.05**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**1**
DC	85.00%	65.00%	70.00%	65.00%	70.00%	50.00%	50.00%
JK	50.00%	50.00%	60.00%	60.00%	50.00%	55.00%	55.00%
JE	60.00%	60.00%	65.00%	60.00%	45.00%	55.00%	60.00%
KL	50.00%	60.00%	15.00%	30.00%	45.00%	45.00%	35.00%
Average	**61.25 %**	**58.75 %**	**52.50 %**	**53.75 %**	**52.50 %**	**51.25 %**	**50.00 %**

Table 1 presents classification accuracy using MFCC feature with HMM as classifier. The aim of this step is to set the optimal frame length to be used. Results show that the best accuracy is obtained with a frame length of 50ms with an average recognition rate of 61.25%.

In the following experiments samples will be segmented into frames of 50ms each with 50% overlap.

Table 2 shows the classification results for the four speakers obtained from SVM OAA method applied to MFCC features. These results were run using polynomial and Gaussian kernel. The classifier gives accuracy for testing data are in the range of [39% 53%] with an average of 45.83% for the polynomial Kernel, and in the range of [45% 67%] with an average of 54.58% for Gaussian kernel.

TABLE II. CLASSIFICATION ACCURACY USING SVM/OAA AND MFCC

	Testing	**Training**	**Kernel**	**C**	**σ**	**SV**
DC	49.17%	100.00%	poly	1	5	66
	66.67%	99.58%	gaussian	11	11	203
JE	39.17%	100.00%	poly	11	3	79
	55.83%	97.50%	gaussian	1	10	211
JK	53.33%	100.00%	poly	1	6	81
	50.00%	100.00%	gaussian	11	9	218
KL	41.67%	100.00%	poly	1	9	88
	45.83%	96.67%	gaussian	1	10	220
AVERAGE	**45.83%**	**100.00%**	**poly**			
	54.58%	**98.44%**	**gaussian**			

TABLE III. CLASSIFICATION ACCURACY USING SVM/OAO AND MFCC

	Testing	**Training**	**Kernel**	**C**	**σ**	**SV**
DC	65.00%	97.91%	poly	1	10	110
	69.17%	95.42%	gaussian	21	91	72
JE	45.00%	100.00%	poly	11	2	45
	52.50%	100.00%	gaussian	21	10	108
JK	53.33%	100.00%	poly	1	5	53
	55.00%	100.00%	gaussian	21	10	112
KL	46.67%	100.00%	poly	1	9	60
	43.33%	99.17%	gaussian	11	10	114
AVERAGE	**52.50%**	**99.48%**	**poly**			
	55.00%	**98.65%**	**gaussian**			

The same data sets with same features applied to a SVM OAO are presented in table 3 which shows a testing data average classification rate of 52.50% for polynomial kernel with a minimum of 45% and a maximum of 65%, and 55% for Gaussian with classification rate between 43% and 69%.

We can remark that for both methods Gaussian kernel gives better results, and that One Against One method of Support vector machine are slightly better than One Against All.

TABLE IV. CLASSIFICATION ACCURACY USING SVM/OAA AND LFCC

	Testing	**Training**	**Kernel**	**C**	**σ**	**SV**
DC	52.50%	100.00%	poly	1	10	79
	51.67%	99.58%	gaussian	11	5	202
JE	35.00%	53.33%	poly	1	1	159
	40.83%	93.33%	gaussian	1	7	198
JK	45.83%	49.58%	poly	21	1	234
	50.00%	84.16%	gaussian	1	10	203
KL	39.16%	100.00%	poly	1	7	112
	36.67%	88.75%	gaussian	1	10	210
AVERAGE	**43.12%**	**75.73%**	**poly**			
	44.79%	**91.46%**	**gaussian**			

We can remark from table 4 that both kernels used gives an important range of classification rate between minimum and maximum figures giving a difference of 17.5% for polynomial and 15% for Gaussian. The results of Gaussian kernel are better in term of accuracy than polynomial kernel.

Table 4 presents results is obtained from applying LFCC feature to OAA strategy of Support Vector Machine with same kernels used for MFCC.

TABLE V. CLASSIFICATION ACCURACY USING SVM/OAO AND LFCC

	Testing	Training	Kernel	C	σ	SV
DC	46.67%	100.00%	poly	41	9	46
	51.67%	99.16%	gaussian	21	10	90
JE	35.00%	62.50%	poly	1	1	76
	45.83%	96.25%	gaussian	21	8	92
JK	42.50%	53.33%	poly	11	1	98
	50.00%	81.67%	gaussian	1	8	120
KL	39.16%	100.00%	poly	1	8	61
	40.00%	100.00%	gaussian	41	5	110
AVERAGE	**40.83%**	**78.96%**	**poly**			
	46.88%	**94.27%**	**gaussian**			

The same experiment conducted using One Against One strategy is described in table 5 which shows that for all speakers Gaussian Kernel gives better results with an average accuracy for testing data of 46.88%.

Among the four result tables using SVM as classifier with its two strategies, we can remark that:

- very similar trends can be observed: the best performance is achieved by Gaussian kernel:
- for almost results with polynomial kernel, One against one strategy is better than One Against All:
- MFCC describes better emotional trends in speech signal.

TABLE VI. CLASSIFICATION ACCURACY USING HMM

	MFCC	LFCC
DC	85.00%	50.00%
JE	60.00%	45.00%
JK	50.00%	60.00%
KL	50.00%	25.00%
Average	61.25%	45.00%

Table 6 reflects the classification rates obtained from HMM classification method applied to MFCC and LFCC features. In this table, it is obvious that MFCC performs better than LFCC when for three of the four speakers MFCC feature, and in average, the first method gives better recognition rates.

The results demonstrate that HMM is better classifier than SVM with its two strategies with an average accuracy of 61.25%. One reason for this might be that HMM can model dynamic changes of acoustic features in given emotional state. Moreover, MFCC proves to be most descriptive than LFCC, the results obtained using this feature is steadily better than those from LFCC features.

The classification dispersion between speakers reaching 35% can be presented as the field in which improvement can be made. This may be due to the fact that it is difficult even for human subjects to determine the emotion of some recorded utterance.

VI. CONCLUSION

In this work, a classification methods using SVM and HMM was designed by empirical guidance. These methods were applied to SAVEE data base using a set of features including MFCC and LFCC.

Experiment results demonstrate that our method can serve as a viable approach for the classification of emotions from speech with a recognition rate reaching 61.25%. Besides, we have been able to conclude that MFCC describes better emotional state in speech than LFCC, and that HMM is better classifier than SVM for the used set of data.

Many future modifications can be integrated within this framework. We can for instance develop the used methods with a larger emotional speech databases with reasonably large numbers of speakers in order to improve the generalization of the classification results.

REFERENCES

[1] D.Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '04,vol. 1, pp. 593-596, Montreal, May 2004.

[2] J. Kreiman and B. R. Gerrat, "Perception of aperiodicity in pathological voice", Acoustical Society of America, vol.117, pp. 2201-2211, 2005.

[3] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, February 1975.

[4] L. Rabiner, R. Schafer, "Digital Processing of Speech Signals"",Pearson Education, 1978.

[5] R. Le Bouquin, Enhancement of noisy speech signals: application to mobile radio communications", Speech Commun., 18 (1) (1996), pp. 3–19.

[6] Specht, D. F., "Probabilistic neural networks for classification, mapping or associative memory", Proceedings of IEEE International conference on Neural Network, Vol. 1, pp.525-532, Jun. 1988.

[7] V. Petrushin, Emotion recognition in speech signal: experimental study, development and application, in: Proceedings of the ICSLP 2000, 2000, pp. 222–225

[8] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the ICASSP 2004, vol. 1, 2004, pp. 577–580.

[9] M. Slaney, G. McRoberts, Babyears: a recognition system for affective vocalizations Speech Commun., 39 (2003), pp. 367–384

[10] Inanoglu, Z., & Caneel, R. (2005). Emotive alert: Hmm-based motion detection in voicemail messages. In Proceedings of the 10th international conference on intelligent user interfaces (IUI'05), no. 585. San Diego, California, USA: ACM Press.

[11] S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification", In Proc AVSP, pp. 185-190, 2008.

[12] X.Zhou, D.G.Romero, R.Duraiswami, C.E.Wilson, S.Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.559-564,2011.

[13] S.B. Davis, P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics speech and Signal Processing, Vol ASSP-28, No 4, pp 357-366,August 1980.

[14] L.R Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition",Proceedings of the IEEE, Vol 77, No 2, February 1989.

[15] V.N.Vapnik V.N, "Statistical Learning Theory", Wiley-Interscience, New York, 1998.

[16] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. C. Woodland, The HTK Book, version 3.4.2006.

[17] S. Canu, Y. Grandvalet, V. Guigue, A. Rakotomamonjy, SVM and Kernel Methods Matlab Toolbox, Perception Systèmes et Information, INSA de Rouen, France,2008.