RECOGNIZING EMOTIONS FOR THE AUDIO-VISUAL DOCUMENT INDEXING

Xuan Hung LE, Georges QUÉNOT, Eric CASTELLI
Laboratoire CLIPS-IMAG – 385 rue de la Bibliothèque
BP. 53 – 38041 Grenoble Cedex 9 – France
(lex@imag.fr, Georges.Quenot@imag.fr, Eric.Castelli@imag.fr)

Abstract

In this paper, we proposed using MFCC coefficients and a simple but efficient classifying method: Vector Quantification (VQ) to perform speaker-dependent emotion recognition. Many other features: energy, pitch, zero crossing, phonetic rate, LPC… and their derivatives are also tested and combined with MFCC coefficients in order to find the best combination. Other models, GMM and HMM (Discrete and Continuous Hidden Markov Model), are studied as well in the hope that the use of continuous distribution and the temporal evolution of this set of features will improve the quality of emotion recognition. The accuracy recognizing five different emotions exceeds 80% by using only MFCC coefficients with VQ model. This is a simple but efficient approach, the result is even better than that obtained with the same database in evaluation by humans with the listening and judge no compare [8].

1.Introduction

Dealing with the speaker's emotion is one of the latest challenges in speech technologies. It plays particularly an important role in speech synthesis as well as in speech recognition. In this latter case, emotion recognition's objective is to determine the emotional state of the speaker out of the speech samples. Its possible applications include from help to psychiatric diagnosis to intelligent toys and games, and becomes recently and rapidly subject of researches. Our emotion recognition system aims at indexing the content of video documents for the purpose of content-based retrieval.

This paper describes firstly some features used in emotion recognition. Secondly, some approaches are briefly presented. And finally, it presents our experimentations with the deferent features, with Vector Quantification Model, Gaussian Mixture Model and Hidden Markov Model (HMM).

In order to carry out the experimentations, we have used a part of the Danish Emotional Speech corpus of Aalborg University [8]: the 13 different sentences uttered in the five styles of emotion: Happy, Angry, Surprise, Neutral, Sad. The four actors and actress are used for the role of speakers. This leads to a total of 13 x 5 x 4 = 260 sentences, which were divided in two non-overlapping sets by random selection: one with 175 sentences for training purposes, and another with 85 sentences for test.

2.Features for Emotion Recognition

A crucial problem of all recognition systems is the selection of the set of features to be used. Pitch and Energy are classical features that are used in a majority of applications and research systems [9]. Many other measures relating to pitch, to energy, to durations, to tunes, to spectral, to intensity…have also been studied. Sinéad McGilloway introduced up to 32 coefficients [4], Pierre-Yves Oudeyer used 20 features [5]. However, the set of the most efficient features for emotion recognition is still vague.

The use of MFCC coefficients in emotion recognition is nowadays quite sparse. For our approach, MFCC coefficients (Mel Frequency Cepstral Coefficients) are primarily used. To improve the quality of the system, other features will be combined with these MFCCs.

Output of feature extraction process is a series of feature vectors that corresponds to consecutive frames of speech signal. The dimension of these vectors indicates the numbers of features used.

2.1 Mel Frequency Cepstral Coefficients

The MFCC coefficients approach was the second improvement over the direct use of the Linear Prediction parameters. This is a feature derived based on the psychoacoustics modelling which studies human auditory perception. The 'Mel' is a unit of measure of perceived pitch or frequency of a tone. The Mel Frequency Cepstral Coefficients are calculated by applying Discrete Cosine Transform (DCT) of the Mel-scaled log filter bank energies. If f_H and f_M are frequency and the Mel scales respectively, then the approximation of the Mel scale with respect to the frequency scale is given by:

$$f_M = x * \log_0\left(1 + \frac{f_H}{y}\right)$$

where x = 2595, y = 700 being the most commonly used values nowadays.

We used 24 or 64 triangular mel-filter banks - that mean at most 24 or 64 coefficients of MFCC can be obtained. In our experimentations, the number of MFCCs and their derivatives is varied to find the best configuration.

Figure 1 : construction Mel Frequency Cepstral coeffcients

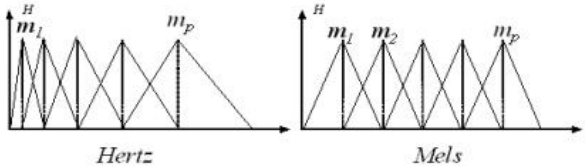

Figure 2 : Triangular Mel-filter bank.

2.2 Energy features

When introducing instantaneous values of energy in the feature vectors with MFCC coefficients, in some model, for example in VQ model, too high values of energy will affect the role of MFCCs, and conversely: with too low values of energy, information of energy becomes useless. So, we balanced it by the normalization of the logarithm of the instantaneous values of energy in each frame between [MIN, MAX]. We used here 1.0 for MAX and MIN gets the negative of mean absolute value of MFCC. The derivative of logarithm of energy is also used like a feature coefficient.

2.3 Pitch features

Pitch and its derivatives are features that are almost always used [9][10]. It seems to be also a good measure that reflects considerably the emotional state of the speaker. Experimentations of Albino Nogueiras [9] have shown that by using information of pitch 75% correct rate can be reached with their corpus.

For our experimentations, MFCC coefficients show always the best, so we just tried to combine the instantaneous information of pitch with MFCC in the objective of improving the quality of recognition.

In order to extract the pitch, we performed a simple auto-correlation at every frame. After thresholding, the position of the first maximum is determined and this raw value is used like instantaneous value of pitch. The other types of information of pitch can be obtained from these values.

Because the global pitch depends heavily on the speaker's nature, in order to combine the pitch with MFCC coefficients, the values of pitch must be also normalized.

2.4 Other features

Zero crossing features, phonetic rate features, LPC features (Linear Prediction Coding) are the other features that we have experimented. Zero crossing features present a similar behaviour as pitch ones. Phonetic rate features are determined from the number of samples occupied by phonemes. These features must be also normalized when combining in the feature vectors.

3. Approaches in Emotion Recognition

In recent years, a lot of approaches have been performed with different set of features and different models: Linear Discriminant [4], K Nearest Neighbors [3], Neural Network [3], Support Vector Machine [4] …; and Hidden Markov Model seem to be the best with about 70% - 80% correct [6][9], indeed, this type of model allows the system to capture efficiently temporal behaviours of features.

Our proposal consists in studying the performance of MFCC features for emotion recognition while modelling these features using several types of model.

In the first approach, we used Vector Quantification model. With this model, the temporal evolutions of features are ignored. Indeed, each emotional model is only represented by a set of specific vectors (called centroids) in any order. Specific vectors are constructed by means of training the models, so the different emotional models have different sets of specific vectors. In fact, each specific vector is the mean vector of a group of similar feature vectors.

In the model training process, feature vectors in the training set is divided and gathered into N groups by K-means algorithm (N centroids). N is the primary parameter of this type of model and must be chosen appropriately. If N is too small or too high, the degree of specificity obtained for each emotion represented by the set of specific vectors will be attenuated.

In the recognition stage, a test utterance will be evaluated with all emotion models. The model that gives the smallest distance will be chosen.

Distance between a utterance X characterized by a series of feature vectors: $\{x_1,...,x_M\}$ and a model λ represented by a set of specific vectors $\{y_1,...,y_N\}$ is calculated by following formula:

$$D=\sum_{m=1}^{M}\min(d(x_m,y_n) \mid n=1..N)$$

Discrete Hidden Markov Model (DHMM) is a combination of vector quantification and Markov process. Thanks to Markov process, temporal evolutions of features is captured and presented in the form of probability distribution. Besides the number of centroids, the number of states and the relation between these states are also important parameters. For our experimentation, the number of states was varied from 1 to 64 and, by changing the relation between states, we can obtain estimations on the relative performance of the ergodic models and left-right models in emotion recognition.

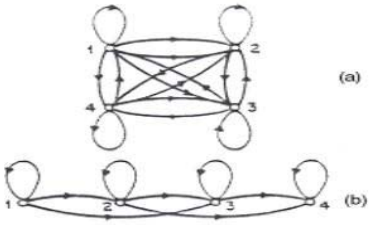

Figure 3 : a) Ergodic Model of 4 states
b) Left-Right Model of 4 states

With this approach (using DHMM), the temporal characteristics of features in emotion recognition are already exploited. However, we can expect that, the use of continuous distribution (normally Gaussian distribution) in place of vector quantification give better performance, Continuous Hidden Markov Model (CHMM) is the perfect combination of these two aspect: Markov Process and Continuous distribution.

$b_j(O_t)=\sum_{m=1}^{M} c_{jm}.\eta$ is general representation of probability density function of series of observations O_t where c_{jm} is a gain coefficient, η is gaussian distribution.

Gaussian Mixture Model may be considered like a particular case of CHMM whose number of states is equal to 1.

So, the last important parameter of two models CHMM and GMM is number of gaussian mixtures: M.

4. Experimental Framework

4.1 Danish Emotional Speech Database (DES)[8]

Danish Emotional Speech Database was recorded at the Center for PersonKomunikation (CPK), Aalborg University, Denmark as a part of the VAESS project. The design of this database is especially oriented toward speech synthesis purposes, but it can also provide a first approximation to emotional speech analysis and emotion recognition. DES was recorded in an acoustically damped sound studio at Aarhus theatre. A high quality microphone was used, which did not influence the spectral amplitude or phase characteristics of the speech signal [8].

Four actors familiar with radio theatre were employed for the recording of DES (two men and two women). The utterances were sampled with 16 bits at 20 KHz. Three differences kinds were recorded: single words, sentences and passages of fluent speech and consist of different kinds of sentence: affirmative, exclamatory and interrogative sentences.

A listening test (listening and judges with no compare) was performed on DES at CPK with 20 normal-hearing listeners (10 of each gender) obtained 67% correct on average [8].

In our first experimentation, speaker-dependent emotion recognition, 260 sentences of corpus are randomly divided into 2 sets: training (175 sentences, 70%) and test (85 sentences, 30%) in such a way that the quantity of sentences is balanced across the training and test parts of corpus for each person and each emotion.

In our second experimentation, speaker-independent and sex-independent emotion recognition, four speakers in corpus are divided into 2 parts:

+3 for training and 1 for test

+2 men (women) for training and 2 women (men) for test.

4.2 Emotion Recognition Results

4.2.1 Influence of features

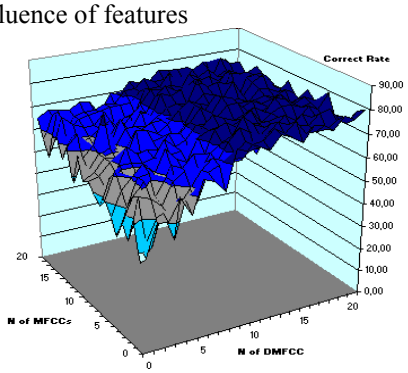

Figure 4 : Influence of MFCC features

Figure 4 shows the results obtained (from our speaker dependant test) when changing the number of MFCC and number of derivatives of MFCC (DMFCC) from 0 to 20 independently. So, these results are given for 21 x 21 - 1 = 440 combinations which are tested with Continuous Hidden Markov Model 4 states–7 mixtures. From these results we found that the efficient combination between MFCC and DMFCC locate on the region from 10 MFCC and influence of DMFCC to the quality of system is really weak on this region. In other words, from these results we can notice that temporal variation of MFCC does not significantly reflect the emotional states of speakers.

We chose anyway 12MFCC + 12 DMFCC as an optimal combination because they reside on the optimal region and furthermore, this is the combination most commonly used in speech recognition.

Table 5 gives the correct rates of isolated features. It may be that the other configurations of CHMM will give better performances to some of these isolated features. But in comparison with performance of MFCC 76%, we found that the selection of MFCC to play the role of primary features is reasonable.

Features	Correct Rate
Energy	36
Energy + Derivative of Energy	48
Pitch	35
Pitch + Derivative of pitch	41
Zero crossing	43
Zero crossing and Derivative	43
Phonetic rate	36
12 MFCC + 12 DMFCC	76

Table 5 : Correct rate of isolated features.

From our result, the influence of other features when combined with MFCC is unclear. For this configuration it may be better but for other configuration it's worse. It can be explained by the size, which is not so big, of the used database. Therefore, we choose always 12 MFCC + 12 DMFCC for all the other tests.

4.2.2 Influence of model parameters

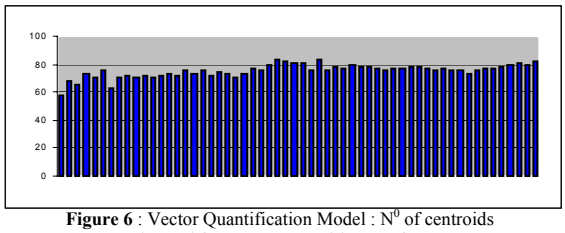

Figure 6 : Vector Quantification Model : N^0 of centroids changed from 10 to 512 with step of 10

With 12 MFCC + 12 DMFCC applied in VQ Model and by changing number of centroids. From the results, we found that in general this model gives quite good results, from the position of 160 to 200 centroids, the best results (>80%) were reached. This is already quite high value. That proclaims not only the performance of MFCC features, but also the efficiency of simple VQ model.

Choosing 170 (in the optimal region) for the number of centroids for Discrete Hidden Markov Model, Table 7 shows the results obtained when changing the number of states in DHMM:

Features	Number of states						
	1	2	4	8	16	32	64
12 MFCC + 12 DMFCC	58,82	56,47	56,47	56,47	58,82	49,41	57

Table 7 : Results obtained with Discrete Hidden Markov Model

Clearly, these results do not satisfy our expectation. However, it is explicable: in VQ model we compare directly feature vectors with model's specific vectors without taking care of both their distribution and their

order. On the contrary, this information of the distribution and the information of the order are crucial for both training and recognition process in Markov model. So, it's normal when the results of DHMM are not good as expected.

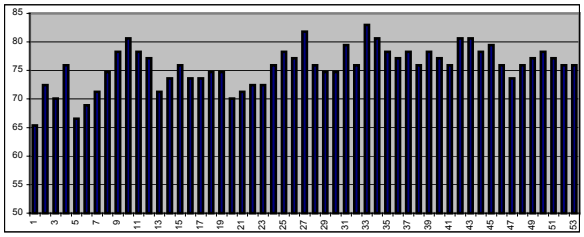

Figure 8 : Result obtained with GMM model

Gaussian Mixture Model models features by gaussian continuous distributions. By doing so, it can avoid the information degradation caused by vector quantification process. Figure 8 show the results obtained with GMM by increasing the number of mixtures. Globally, the more the number of mixture increases, the better the results are. In general, this is a promising approach.

In order to compare with results obtained from GMM model, we experimented firstly CHMM 1 state; their number of mixtures is changed from 1 to 50. Although there are several differences in the shape of results, both show similar performance.

Figure 9 contains results obtained from CHMM when changing the two parameters: number of states and number of mixtures. In general, with GMM or with CHMM, about 80% is the maximum value we can reach. 70% is the average value.

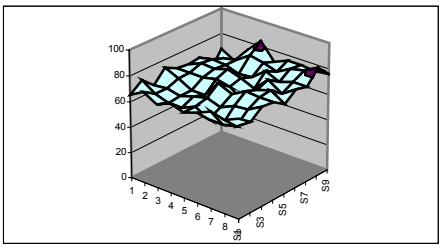

Figure 9 : Result distribution of CHMM model when changing number of states and number of mixtures from 1 to 10.

Our results also indicated that ergodic models show always better in comparison with left-right model when working with MFCC features in the domain of emotion recognition. It corresponds to notices of Tin Lay Nwe[6] as well.

In the case of speaker-independent or sex-independent emotion recognition, all obtained results are normally from 30% to 40%, particularly case is about 50%-55%. In the future, with new and bigger corpora, we expect to significantly improve these figures.

CONCLUSION

Four emotion recognition approaches have been evaluated, mainly in combination with a set of MFCC features. Except for DHMM model; the three other models have almost the same performance. The maximum recognition rate for the speaker dependant system is about 80% and 70% on average is a typical rate. Especially, the very simple technique of the Vector Quantification model gives us the highest efficiency and the highest stability in emotion recognition.

From the results of figure 6, figure 8, figure 9 and table 7, in a certain degree, we can conclude that, temporal characteristics of MFCC features do not play a role important in emotion recognition. It is clearly proven by the results of two models VQ and GMM: although these two models do not capture temporal behaviours, their performances are not worse than performances of the other models DHMM and CHMM.

According to [14], by using only energy and pitch information, their system has already reached a similar performance. Trying to combine the information of different features in order to improve the quality of emotion recognition system is still a challenge for our future work.

References

[1] K. R. Scherer, T. Johnstone, J. Sangsue - L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole - Université de Genève

[2] Frank Dellaert, Thomas Polzin and Alex Waibel- Recognizing emotion in speech - School of Computer Science Carnegie Mellon University Pittsburgh, Pennsylvania 15213-3890

[3] Valery A. Petrushin – Emotion in speech: Recognition and Application to call centers - Proceedings of the 1999 Conference on Artificial Neural Networks in Engineering (ANNIE '99)

[4] Sinéad McGilloway, Roddy Cowie, Ellen Douglas-Cowie, Stan Gielen, Machiel Westerdijk and Sybert Stroeve – Approaching automatic recognition of emotion from voice: A rough benchmark - National University of Ireland, Maynooth Queen's University of Belfast, Northern Ireland University of Nijmegen

[5] Pierre-yves Oudeyer - Novel Features and Algorithms for the Recognition of Emotions in Human Speech - Sony Computer Science Lab, Paris, France.

[6] Tin Lay Nwe, Foo Say Wei, Liyanage C De Silva - Speech Based Emotion Classification - Electrical and Electronic Technology, 2001. TENCON. Proceedings of IEEE Region 10 International Conference on Volume: 1, 2001 Page(s) : 297 -301 vol.1

[7] Douglas A. Reynolds - A Gaussian Mixture Modeling Approch to Text-Independent Speaker Identification - Thesis - Georgia Institute of Technology 1992.

[8] Inger Samsø Engberg & Anya Varnich Hansen - Documentation of the Danish Emotional Speech Database DES - Center for PersonKommunikation Department of Communication Technology Institute of Electronic Systems Aalborg University – Denmark

[9] Speech Emotion Recognition Using Hidden Markov Models Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño - Research Center TALP, Universitat Politècnica de Catalunya. SPAIN. Eurospeech 2001

[10] INTERFACE Project, "Multimodal analysis/synthesis system for human interaction to virtual and augmented environments," EC IST-1999-No 10036, coor. F. Lavagetto, 2000–2002, http://www.ist-interface.org.