Proceedings of Engineering & Technology (PET)
pp. 122-126
2nd International Conference on Automation,Control, Engineering and Computer Science (ACECS-2015) Proceedings of Engineering Technology (PET)

Speaker Recognition in Emotional Context

Asma Mansour
Signal,Image and Information Technology Laboratory
National School of Engineering Tunis
Tunis, Tunisia

Zied Lachiri
Signal,Image and Information Technology Laboratory
Physic and Instrumentation Department
National Institute of Applied Science and Technology
Tunis, Tunisia

***Abstract*—This paper attempts speaker recognition in emotional context by developing different descriptors to represent speaker specific emotional information. Performance can be decreased when emotions alter the human voice. So, Mel-frequency cepstral coefficients (MFCC) are used and combined with Energy, Pitch and their first derivations to enhance accuracy. Moreover, we have developed other cepstral features such Linear Predictive Cepstral Coefficients(LPCC) and Formant based on Linear Prediction Coefficients (FBLPC). His fusion with MFCC showed a good accuracy of speaker recognition from emotional speech. HMM is used as a classifier and four different emotional states are considered in experiment : neutral, happy, angry, and sad. Cepstral features yield better accuracy of speaker recognition in emotional talking environment.**

***Index Terms*—Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Formant based on LPC(FBLC), Hidden Markov Model (HMM), Emotional context, Speaker recognition.**

I. Introduction

Speaker recognition in emotional talking environment become one of the most important research area for the field of Human-Machine Interaction[1]. Emotional states influence directly speaker recognition system because every person has different emotions when talking, which are speaker dependent characteristics. As a result, speaker recognition in emotional environment can get many applications such as criminal investigation[2].

In recent years, many studies focused on speaker recognition field in emotional context. Bao et al[3] were concentrated on emotion attribute projection for speaker recognition on emotional speech. They proposed two methods to ease the emotion effects on speaker recognition in emotional context. The first one is the emotion compensation method called Emotion Attribute Projection (EAP).The second method is the linear fusion of two subsystems, the Gaussian Mixture Model-Universal Background Model (GMM-UBM) based system and the Support Vector Machine (SVM) with EAP system. The two approach have been proved successful.Ghiurcau et al [4] proved in first step the important influence of the emotional state upon text independent speaker identification.Then, they used MFCC for describing feature of speech signal and SVM for training the speaker models and testing the system. In two of his recent studies, I.Shahin[5][6]focused on recognition unknown speaker in emotional talking environment.In the first study[5], he proposed two approachs based on both Hidden Markov Models (HMMs) and Suprasegmental Hidden Markov Models (SPHMMs) as classifiers. Speaker identification performance obtained is 79.92% from a set of fifty speakers and six different emotions.In the second study, he proposed a new approach that is based on identifying the unknown speaker using both his/her gender and emotion cues using Hidden Markov Models (HMMs) as classifiers in order to enhance the degraded performance of text-independent speaker identification in emotional talking environment[6]. J. Sirisha Devi et al[7] presented a new approach to identify speaker from emotional speech. This approach enhance in first part the detection of the emotion of the speaker prior using the hybrid Feed Forward Back Propagation Neural Network(FFBN)and Gaussian Mixture Model(GMM) methods. In second part, after recognition emotion by previous hybrid method, HMM are used to recognize speaker.They showed that speaker recognition rate using hybrid method FFBN/GMM in emotion recognition is better than using GMM method.

The purpose of this work is to develop and test speaker recognition performance under emotional states using combination of various descriptors based on two types of features : Continuous features as Energy and Pitch, and cepstral features as MFCC, LPCC and FBLPC.

To evaluate our work , we have applied Hidden Markov Models (HMM) classifier to IEMOCAP database[8].

The organization of the paper is as follows : the next section is committed to describe different feature extraction techniques using the continuous and cepstral features. In the third section is reserved to evaluate experiments, so HMM classifer and IEMOCAP data base will be introduced. Show and interpret results is detailed in section IV. Finally, concluding remarks and possible future work are given.

II. Feature Extraction

The main task of feature extraction procedure is to extract the most relevant characteristics for emotional speech and represent them in feature vector. Two types of features are distinguished : Continuous features(energy,pitch) and cepstral features(MFCC, LPCC, FBLPC).

ISSN: 2356-5608

A. Continuous Features

1) Energy: In nature, energy associated with speech is time varying. Hence, the interest for any automatic processing of speech is to know how the energy is varying with time and specially with short term region of speech. Thus, energy is one of the most important features of speech signal. To get energy value in each frame, the following short term function is used :

$$E(n_o, n_1) = \Sigma_{n_o}^{n_1} |x(n)^2|. \quad (1)$$

2) Pitch: It is an important attribute of voiced speech because it is heavily influenced by the speaker's nature and its global evolution along the utterance. Get the value of pitch frequency in each speech frame, and obtain the statistics of pitch in the whole speech sample. There are many methods to track pitch from speech signal as autocorrelation method and cepstrum method.

B. Cepstral Features

1) MFCC: Any recognition system needs a robust acoustic-feature-extraction technique as a front-end block followed by an efficient modeling scheme for generalized representation of these features. The Mel-frequency cepstral coefficients(MFCC) are short term spectral features based on the characteristics of the human ear's hearing. Its represent the phonetic content of speech signal. In fact, in a word each tone has a frequency measured in Hz. MFCC use a nonlinear Mel scale frequency to simulate the human auditory system. Mel scale is linear below 1000 Hz and logarithmic spacing above 1000 Hz. The conversion from linear frequency f to mel frequency F_{mel} is computed by the following approximate formula :

$$F_{mel} = 2595 \log_{10}(1 + \frac{f}{700}) \quad (2)$$

Input signal is processed frame by frame. The size of each frame depends on the sampling frequency. The time length is generally in the range of 20-40ms.
The following steps performed to obtain the MFCCs from speech signal :

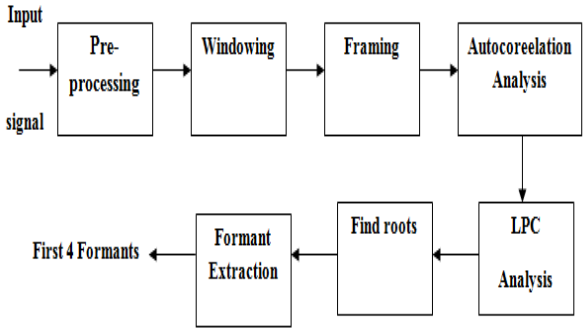
⋆Compute Fast Fourier Transform analysis in each frame.
⋆Get powers of the spectrum using triangular filter banks which are computed according to the mel scale.
⋆Compute the logs of each mel frequencies.
⋆Calculate DCT of all mel log powers.
⋆Get the MFCC coefficients from the amplitudes of resulting spectrum.

2) LPCC: The linear predictor coefficients(LPC) are rarely used as features but they are transformed into the more robust Linear Predictive Cepstral Coefficients (LPCC) features. LPCC includes the characteristics of particular channel of speech, and the same person with different emotional speech will have different channel characteristics. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC), which is according to the all-pole model. However, unlike MFCC which are based on perceptual frequency scale, such as Mel-frequency scale , the LPCC are founded on Perceptual Linear Predictive (PLP) analysis.

3) FBLPC: Formants are the resonance frequency of the vocal tract. So, it characterize the evolution of vocal tract during the speech which is a difficult task because of his coupling with the nasal cavities. Thus, the question is how to interpolate the transfer function of the vocal tract from the speech spectrum ?. Formant frequency change from a person to an other and it has more energy than any other frequency which is good for speaker recognition. This approach propose, for obtaining formants, to compute firstly LPC coefficients for each frame using autocorrelation method Levinson-Durbin Recursion algorithm[9]. Then, based on the above coefficients, formant frequencies are computed by solving the root in the LPC all-pole polynomial. Obviously, the roots have a real, imaginary parts and the phase spectrum is displayed to get the formant frequency [10]. The features of the pole are solved using :

$$F = \frac{f_s \theta_o}{2\Pi} Hz. \quad (3)$$

where H(z) is the transfer function, $z = r_o e^{\pm j\theta o}$ is the complex root pair of the LPC polynomial, F is the formant and f_s is the sampling frequency. Formant frequencies of all the frames of a speech sample are computed. First four estimated formants were considered. The flowing diagram shows different steps to get FBLPC :



(a)

FIGURE 1: FBLPC Feature Extraction.

III. EXPERIMENTAL EVALUATION

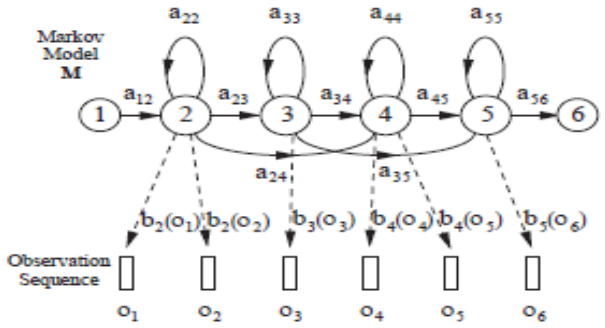
In this paper , we performed speaker recognition in emotional context using two kind of features : Continuous and cepstral features.

A. Classification System

Hidden Markov Models (HMM) is famous classification technique in the field of speaker and speech recognition[11] as well as in speaker recognition in emotional environment[6]. Let $O = (o_1 o_2 o_3 ... o_n)$ observation vector, N is the total

number of states, and $X = (X_1, X_2, \ldots, X_T)$ state of observation at time t giving the observation O_t.

In the training phase, one HMM for each emotional speech is obtained to mean that the parameters of HMM model are estimated using training feature vectors. Each HMM model can be characterized by the following parameters[12] : the number of states N, the initial probability or probability of being in state i at time 0 $p_i = P(q_0 = i)$; the state transition probability $a_{ij} = P(q_t = j|q_{t-1} = i)$, observation probability density b_i which is the probability of observing o_t in state i.



(a)

FIGURE 2: HMM model.

HMM is a statistical approach for modeling generative sequences characterized by a set of observable sequences. So, it requires large amount of training data for effective estimate of the model parameters.In our case, 70% of the samples were used as training set and the rest considered as a test set. Firstly, we have evaluated the topology of the HMM by varying the number of states and the number of mixture components per state for various feature sets. Bakis model with one Gaussian mixture is chosen. Training and testing of HMM classifier were accomplished using the Hidden Markov Toolkit(HTK)[13].

B. Data base

To evaluate our current work, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database which was collected for studying multi-modal expressive dyadic interactions.This database contains audio, video and motion-capture recordings of dyadic mixed-gender pairs of actors. There are five sessions in total. The recorded dialogs are either scripts or improvisation of hypothetical scenarios. Dyadic sessions of approximately five minute length were recorded and were manually segmented into utterances.
Each utterance has been evaluated by human annotators categorically over the set of :angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other and dimensionally over the axes of :valence, activation, dominance.
In our experimental studies, we considered 2218 emotional speech data which were collected from 10 speakers (5 males and 5 females) belonging to the four emotional states that we examine : Neutral, happy, angry and sad distributed on five sessions. Table I shows the distribution of the corpus over classes of emotions :

TABLE I :IEMOCAP Corpus utterances four emotions

	Neutral	Happy	Angry	Sad	Total
Male	320	150	338	335	1143
Female	311	173	320	271	1075
Total	631	323	658	606	2218

IV. RESULTS AND DISCUSSION

In our studies, the sampling frequency for all utterances is 16 Khz. The length of frames of speech samples is 800 points with overlap 400 points of frame-stepping.

In first experiment, we have combined the MFCC (13 coefficients) with continuous features : energy and pitch. So, same length of frame and overlap is used for all descriptors in order to combine their feature vectors. Then, the next study was concentrated in combination of MFCC with cepstral features such LPCC and FBLPC.

As is shown at the table II , different features combination results in different recognition accuracy rate used for the 10 actors. To the IEMOCAP Database, the feature combination of Energy and Pitch has the worst recognition rate. That may because these two are simple prosodic features with few numbers of dimensions.

MFCC gives the best results for different emotions. Neutral and sadness emotion gives the highest accuracies respectively 80.22% and 85.86%. Furthermore, combined with continuous features, MFCC improves the accuracy of the energy and pitch which demonstrates the benefit of cepstral features in speaker recognition especially in emotional context.

In the second study, cepstral features were combined with the MFCC. Accuracies's results were good. Firstly, combined with MFCC, LPCC give an accuracy 73.08% for neutral emotion and 81.82% for sad emotion. Then, the feature combination of MFCC+FBLPC, classification accuracies were enhanced compared with feature combination of MFCC+LPCC and best rates stay for neutral and sad emotions respectively 80.77% and 86.36%. As a consequence, the use of FBLPC made to reduce the drawbacks of LPC features that it contains prosodic features as well as spectrum features. The table III illustrates the different results :

TABLE II :Accuracy classification for neutral emotion

		MFCC		Energy		Pitch		Emotion
		*	**	*	**	*	**	
MFCC	*	80.22%		55.50%	51.65%	78.98%	79.67%	Neutral
	**		75.86%	61.64%	53.30%	76.92%	74.18%	
	*	77.32%		53.61%	51.55%	64.95%	65.98%	Happy
	**		59.80%	48.45%	46.39%	57.37%	58.76%	
	*	71.58%		55.26%	52.63%	70.53%	67.89%	Angry
	**		63.86%	53.16%	47.37%	60.53%	61.05%	
	*	85.86%		85.86%	66.67%	85.35%	85.35%	Sad
	**		82.32%	78.28%	66.16%	82.32%	83.33%	

* : Normal
** : First derivation

TABLE III :Classification accuracy of combined spectral features

Emotion	MFCC	MFCC+LPCC	MFCC+FBLPC
Neutral	80.22%	73.08%	80.77%
Happy	77.32%	61.86%	62.89%
Angry	71.58%	70.53%	71.58%
Sad	85.86%	81.82%	86.36%

As a general comment, perfect features for speaker recognition does not exist in any environment yet. But, the compromise which must be realized is between speaker recognition, robustness, and feasibility. Fusion of various types of features often provides additional gains.

V. Conclusion

In this paper, speaker recognition in emotional environment performance is evaluated using the IEMOCAP database. Four emotions are presented in speech of each speaker : Neutral, Happy, Angry and Sad. Hidden Markov Models are used to develop the speaker recognition model using several features. It's concluded that emotions have an important influence on performance of speaker recognition and fusion of cesptral features give the best result but it can be ameliorated with more robust features.

In the future studies, we will try to develop a new kind of descriptors that is more robust in order to ameliorate classification accuracies. Moreover, it is recommended to add visual features in the goal of enhancement of speaker recognition system.

References

[1] R. W. Picard. Affective computing. *MIT Media Lab Perceptual Computing Section Tech*, No. 321, 1995.

[2] S. Furui. Speaker-dependent-feature-extraction, recognition, and processing techniques. *Speech Communication*, vol. 10 :vol. 10, pp. 505–520, Mar. 1991.

[3] Mingxing Xu Huanjun Bao and Thomas Fang Zheng. Emotion attribute projection for speaker recognition on emotional speech. *INTERSPEECH*, 2007.

[4] Corneliu Rusu Marius Vasile Ghiurcau and Jaakko Astola. Speaker recognition in an emotional environment. 2011.

[5] Ismail Shahin. Identifying speakers using their emotion cues. *International Journal of Speech Technology*, Volume 14, 2011.

[6] Ismail Shahin. Speaker identification in emotional talking environments using both gender and emotion cues. *978-1-4673-2821-0/13/*$31.002013IEEE$, 2013.

[7] Y. Srinivas J. Sirisha Devi and Siva Prasad Nandyala. Automatic speech emotion and speaker recognition based on hybrid gmm and ffbnn. *International Journal on Computational Sciences & Applications (IJCSA)*, Vol.4, No.1, February 2014.

[8] C. Lee A.Kazemzadeh E. Mower S. Kim J. Chang S. Lee C. Busso, M. Bulut and S. Narayanan. Iemocap : Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, December 2008.

[9] J. Makhoul. Linear prediction : A tutorial review. *Proc. of IEEE*, vol. 63 :no. 4, pp. 561–580, 1975.

[10] P.Chandra Srivastava, S. Nandi. Formant based linear prediction coefficients for speaker identification. *Signal Processing and Integrated Networks (SPIN), 2014 International Conference*, 20-21 Feb. 2014.

[11] 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.

[12] Rabiner L. R. and Juang B. H. Fundamentals of speech recognition. *Prentice-Hall*, 1993.

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