Speech Emotions Recognition of Joy and Sadness Based on Prosodic and MFCCs parameters

H.Horkous
Laboratory signal and communication Laboratory signal and communication Ecole National Polytechnique Algiers, Algeria
houarihorkous29@yahoo.fr

G.Mhania
Laboratory signal and communication Laboratory signal and communication Ecole National Polytechnique Algiers, Algeria
mhaniag@yahoo.fr

Abstract—This work consists on the automatic recognition of the emotions in the speech, because it plays a very significant role in the communication. The automatic recognition of the emotions potentially had a broad application in the Human Machine Interaction.In this work we use prosodic (pitch, intensity and duration) and cepstral parameters MFCCs (Mel-Frequency Cepstral Coefficients) to analyze the emotions of joy and sadness. These parameters will be used in the automatic recognition of the emotions. The system of recognition is based on the method of classification GMM (Gaussian Mixture Model). The obtained results lead us to observe that the use of the prosodic and MFCCs parameters gives a very acceptable rate of recognition (82.81%).

Keywords—Speech emotion, joy, sadness, prosodic, MFCC, GMM

I. Introduction

The human transmit several and various messages by the voice. Among these messages is the emotion, it has a very significant role in the communication of the human. The interface between the human and the machine will be more comprehensible if the machine recognized the state emotional of the human. The automatic recognition of the emotions potentially had a broad application in the human machine interaction for example: robots, emotion recognition in call center, intelligent tutoring system, In-car board system, diagnostic tool by speech therapists, Telephone banking, computer games,etc...

Therefore, our work consists on the automatic recognition of the emotion of the joy and sadness.

So the goal is to extract the parameters prosodic and MFCC to know the influence of each parameter chosen on the emotions joy and sadness, for exploiting in the emotions recognition system. The extraction of the prosodic parameters is obtained by the Praat software which is free software and easy to use and to interpret. The MFCC parameters are extracting by Matlab software. The second section presents notions on the emotion, the parameters prosodic and the MFCCs parameters. The third section shows the corpus used, the extraction and the analysis of the selected parameters. In the fourth section, the recognition of the two emotions joy and sadness is made with the method of GMM. Finally we finish by a conclusion.

II. Notions On The Emotion, The Parameters Prosodic And The Mfcc Parameters

We present here a short definition of the emotion thus the prosodic parameters and the MFCCs parameters which we chose in order to characterize the target emotions of our work, emotions of the joy type and sadness.

A. Emotion

The fact of expressing an emotion implies a great number of physical and physiological changes (neuronal, activation of certain zones of the brain, increase in the rate of heartbeat, etc). The emotions exist only in reaction to events which they are external or interior. When it is expressed, it can be according to very different means: vocal, gestural, facial, physiological, etc. If only the vocal emotion is considered, that which generates sounds spoken or not, the expression of an emotion passes by physical modifications of the vocal tract and articulation, changes on the level of breathing, saliva or by words or sounds [1].

B. Prosodic parameters

The prosodic parameters make it possible to model the accents, the rhythm, the intonation, the melody of the sentence and are thus very relevant for modeling the emotional state of the speaker [2].

The fundamental frequency F_0or pitch: In speech, the fundamental frequency or pitch characterizes the voiced parts of the speech signal and is related to the feeling height of the voice(Figures 1).The voiced parts have a pseudo-periodic structure and, on these portions, the signal is generally modeled as the sum of a periodic signal T and a white noise. The fundamental frequency is the reverse of the period $T, F_0 = \frac{1}{T}$. The Figure 1 represents the contour of fundamental frequency.

Parameters like the minimum, the maximum, the average, the variance, the range and the standard deviation of pitch are used like significant parameters for the discrimination of the emotions [3], [4], [5], [6].

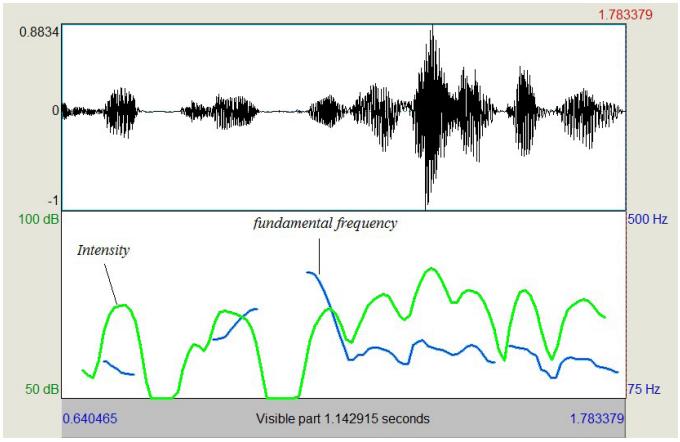


Fig.1. Contours of fundamental frequency and intensity.

Intensity: The intensity corresponds to the variation of the speech signal amplitude caused by a more or less strong energy coming from the diaphragm and causing a variation of the air pressure under the glottis. This descriptor makes it possible to provide a measurement of the sound force of the voice (weak or strong).Pitch, energy, the durations and their derivatives are used for describing the emotional states that are expressed in speech [7].The Figure 1 represents the contour of intensity.

Duration of the voiced trajectory: The third traditional descriptor of the prosody is the rhythm, i.e. rate of the sentence and is measured by the number of vocal units per units of time, for example the number of syllables or phonemes per minute. However, this measurement of the rhythm can be calculated by segmenting the speech signal into syllables. The complex relations between pitch, duration and energy parameters are exploited for detecting the speech emotions[8].

TABLE I. THE PROSODIC PARAMETERS EXTRACTED BY SOFTWARE PRAAT

	Fundamental Frequency (F0 (Hz)) (pitch)				Intensity (dB)				Duration (S)
	Mean	Max	Min	Range	Mean	Max	Min	Range	
Joy	191	354	124	230	70.5	85.3	37.3	48	1.90
Sadness	156	316	95.3	208.6	70.2	84.9	36.7	54.76	3.18

C. *Mel-Frequency Cepstral Coefficients*

MFCCs belong to the family of the cepstral descriptors which base on the representation cepstral of signal. The cepstre has the advantage of allowing a separation of the respective contributions of the source and vocal tract.

MFCCs are obtained while using, for the calculation of the cepstre, a nonlinear frequential scale taking account the characteristics of the human ear, the scale of the Mel frequencies.

The scale of the Mel frequencies is obtained by the following expression [2]:

$$m(f) = 2595 \log\left(1 + \frac{f}{700}\right) \qquad (1)$$

Where f is the frequency in Hertz.

The MFCCs parameters are used in the speech emotion recognition[9], [10],[11].To improve the performance, some recent studies of the speech emotions recognition use the combination enter spectral and prosodic parameters. Information of F0, the log of energy, the formants, energy in Mel and MFCCs are explored to classify the emotions [12].

IV. Extraction And Analyzes Of The Selected Parameters

In this section we present a definition for the corpus chosen thus the extraction and the analysis of the selected parameters.

A. *Corpus*

Berlin Database of Emotional Speech: The Berlin Database of Emotional Speech (Emo-dB) was recorded at the Technical University of Berlin, Germany, and is analyzed very often in speech emotion recognition studies. It contains acted emotional German speech of ten carefully chosen speakers 5 males and 5 females that were asked to pretend six different emotions (anger, joy, sadness, fear, disgust and boredom) as well as a neutral state in ten utterances each of emotionally neutral content. They are characterized by a very high audio quality [13]. Among the speakers who exist in the data base Emo-dB we chose four women and four men and we chose two emotions (joy and sadness) for each speaker.

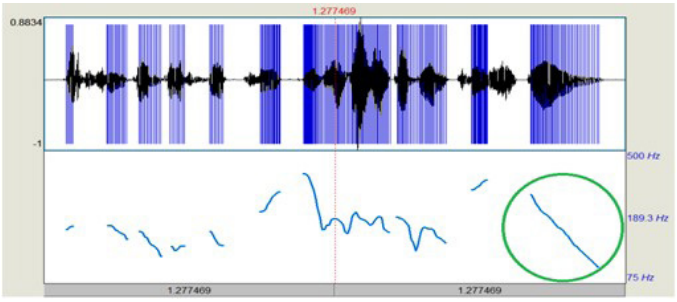
B. *Extraction the chosen parameters*

The results obtained are given in Table 1. This last illustrates the statistics values of the prosodic parameters chosen for the two treated emotions.

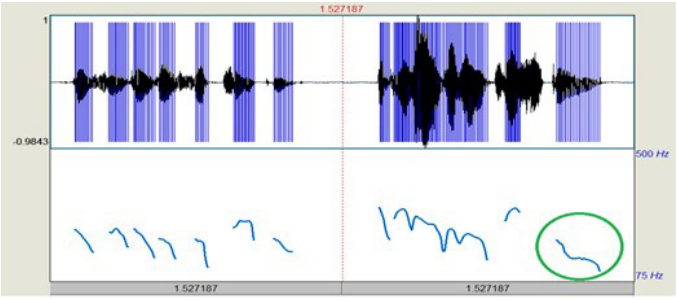
C. *Analyses*

The pitch plays a very significant role, a high pitch correspondent a high frequency sound, a low pitch correspondent a low frequency sound.

We notice according to Table1 that the emotion of joy has a moderate value of pitch mean and we observe a high pitch peak corresponding the same emotion. The sadness emotion has a low value of pitch. The two situations joy and sadness have a broad range of pitch.



a. Emotion of joy



b. Emotion of sadness

Fig.2. Contours of pitch for each emotion

We notice in the Figures 2.a and 1.b during the last word, in the states of joy and sadness the pitch varies in a visible way. But this variation changes from emotion to other.

We remark for the intensity that the joy emotion has a high peak and the state of sadness has a broad range. We notice moderate values of intensity for the two emotions.

It is observed that the duration concerns to the emotion of sadness is very long. And in the state of joy the duration is moderate.

The classifier performance of the emotions is connects with the quality of the data. MFCC is a very powerful technique to analyze the speech signal. Figure 3 illustrates an MFCC form

of two emotions joy and sadness, these results it's obtained by Matlab software.

We have remark on figure 3 that the forms of MFCC for the two emotions (joy and sadness) are vary according to the emotion. This difference which exists between MFCCs allows us used MFCCs parameters in the recognition and classification of speech emotions.

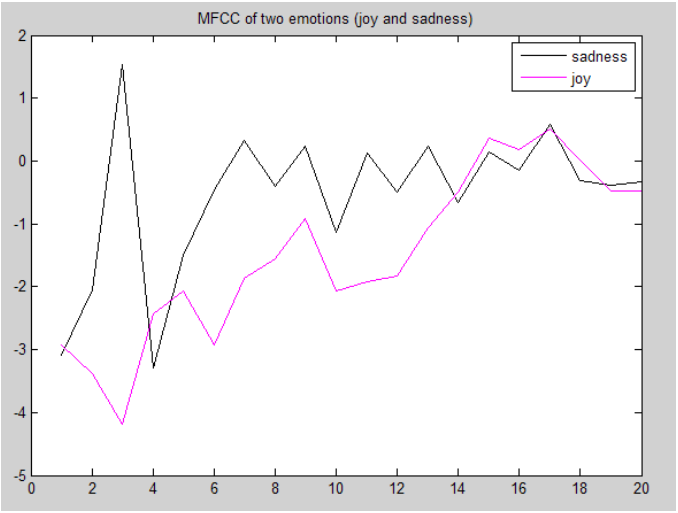


Fig.3. MFCC of two emotions joy and sadness

IV. RESULTS AND EVALUATION

The emotions recognition systems are based on methods of classifications, of which we present a short definition on the method of classification GMM and we describe then the automatic recognition system of the emotional states related to the joy and sadness.

A. Classifier GMM

The GMM consist of modeling, for each classC_q, the data x_d in the form of a sum balanced by the coefficients w_(m,q)of density function of Gaussian probability $p_{m,q}(x)$ [2].

$$p(x/C_q)=\sum_{m=1}^{M} w_{m,q}p_{m,q}(x) \quad (2)$$

With $\sum_{m=1}^{M} w_{m,q} = 1$ for each class Q considered andwhere M is the component count of density considered for the model. Each component is expressed according to its meanμ_(m,q)and of its matrix of covariance $\Sigma_{m,q}$:

$$p_{m,q} = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma_{m,q}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(x-\mu_{m,q})^T(\Sigma_{m,q})^{-1}(x-\mu_{m,q})\right]} (3)$$

The matrix of covariance used is diagonal, i.e. the models are learned by considering the observations associated with each descriptor in an independent way.

For each class, each component of the mixture models a different region from the space of the data called also cluster.

Gaussian mixture model (GMM) was used successfully for the recognition of the emotions [14], GMMs obtain results acceptable and often comparable with other methods of classification [15].

B. The emotions Recognition

The system of recognition is realized by the GMM method in Matlab software. The data used by this system correspond to a representation of the speech signal by the prosodic and the MFCCs parameters.

The system is divided into two parts. In the first part we use only the statistic values of the prosodic parameters. In the second part we use the statistic values of the prosodic parameters and the parameters of MFCCs.

The statistic values of the prosodic parameters used in our system are the mean, the maximum, the minimum and the range of pitch and the range of intensity and the duration. The results are given in tables 2 and 3.

TABLE I. CONFUSION MATRIXES FOR THE SYSTEM OF RECOGNITION, JOY AND SADNESS EMOTION USING ONLY THE PROSODIC PARAMETRES

Emotion	Joy	Sadness
Joy	81.25%	18.75%
Sadness	25%	75%

TABLE II. CONFUSION MATRIX FOR THE SYSTEM OF RECOGNITION, JOY AND SADNESS EMOTION USING THE PROSODIC AND MFCC PARAMETRES

Emotion	Joy	Sadness
Joy	84.37%	15.73%
Sadness	18.75%	81.25%

We notice in tables 2 and 3 that the rates of recognitions are respectively 78.12% and 82.81%. That explains an improvement of performance when we use a combination between the prosodic parameters and the MFCCs parameters.

III. CONCLUSION

In this work we extracted the prosodic parameters (pitch, intensity and duration) and the MFCCs parameters concerning the two emotions (joy and sadness). An analysis is made to know the influence of the extracted parameters on the two emotions (joy and sadness). These parameters are exploited in a recognition system of the two emotions which are indicated.

The obtained results show us that the combination between the prosodic parameters and the MFCCs parameters give recognition rate better than the prosodic parameters only.

REFERENCES

[1] Marie Tahon. Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot, thesedoctorat, paris, 2012.

[2] Chloé CLAVEL. Analyse et reconnaissance des manifestations acoustiques des émotionsde type peur ensituations anormales, Thèse doctorat, paris, 2007.

[3] Schroder, M. (2001). Emotional speech synthesis: A review. In Seventh European conference on speechcommunication and technology, EurospeechAalborg, Denmark, Sept. 2001.

[4] Murray, I. R., & Arnott, J. L. (1995). Implementationand testing of a system for producing emotion by rule in synthetic speech. SpeechCommunication, 16, 369–390.

[5] Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and optimal classificationof speech emotion recognition. In Fourth international conferenceon natural computation (pp. 407–411).

[6] M. Swain, A. Routray, P. Kabisatpathy, J. N. Kundu(2016). Study of prosodic feature extraction for multidialectal Odia speech emotion recognition. IEEE Region 10 Conference (TENCON), 1644-1649.

[7] Cowie, R., & Cornelius, R. R. (2003).Describing the emotional states that are expressed in speech. SpeechCommunication, 40, 5–32.

[8] Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpusbasedspeech synthesis system with emotion. Speech Communication,40, 161–187.

[9] Ververidis, D., &Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48, 1162–1181.

[10] Iliev, A. I., Scordilis, M. S., Papa, J. P., & Falco, A. X. (2010). Spokenemotion recognition through optimum-path forest classification using glottal features. Computer Speech and Language, 24(3), 445–460.

[11] Bitouk, D., Verma, R., &Nenkova, A. (2010, in press). Class-level spectral features for emotion recognition. Speech Communication.

[12] Kwon, O., Chan, K., Hao, J., & Lee, T. (2003). Emotion recognition by speech signals. In Eurospeech, Geneva (pp. 125–128).

[13] Thurid Vogt · Elisabeth André. An Evaluation of EmotionUnits and Feature Types for Real-Time Speech Emotion Recognition. Springer-Verlag (2011) 25:213–223.

[14] B. Schuller, G. Rigoll, et M. Lang. Speech emotionrecognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. Dans Proc. of ICASSP, Montreal, 2004.

[15] A. Batliner, S. Steidl, B. Schuller, et D. Seppi. “you stupid ting box” - children interacting with the aiob robot : a cross-linguistic emotional speech corpus. Dans Proc. Of LREC, pages 171–174, Lisbon, 2004.