ECHAHID HAMMA LAKHDAR UNIVERSITY - EL-OUED
Under the Supervision of the DGRSDT and in collaboration with the CRTI
International PluridisciplinaryPhD Meeting (IPPM'20)
1st Edition, February23-26, 2020
Theme: Modern Technology and Fineness Life

4 à 5 pages

Evaluation the Performance of Different Parameters To Recognize Emotions from Speech

Horkous Houari *, Guerti Mhania**
* Laboratory signal and communication, Ecole National Polytechnique, Algiers, Algeria
** Laboratory signal and communication, Ecole National Polytechnique, Algiers, Algeria
E-mail: houarihorkous29@yahoo.fr, mhaniag@yahoo.fr

Abstract

Recognition emotion in speech signal has attracted much attention and plays an important role in affect computing, artificial intelligence and signal processing areas. The aim of speech emotion recognition system is to extract the information from the speech signal and identify the emotional state of a human being. In this work emotional speech corpus in Algerian Dialect is used for parameters extraction to analyze the emotions of fear, anger and neutral. The selected parameters in our study are the prosodic (pitch, intensity and duration), the unvoiced frames, jitter, shimmer and cepstral parameters MFCCs (*Mel-Frequency Cepstral Coefficients*). The system of recognition is based on the method of classification KNN *(K-Nearest Neighbor).* The obtained results lead us to observe that combined prosodies, jitter, shimmer, unvoiced frames and MFCCs parameters give recognition rate important (84.02%).

Keywords: Algerian Dialect, prosodic, unvoiced frames, MFCC, KNN.

1. Introduction

Emotional speech recognition is an interesting application that is able to recognize different emotional states from speech signal. Speech recognition technology is included in many kinds of advanced study fields. Along with the rapid development of human computer interaction system, emotion in speech is a topic that has received much attention during the last few years. 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...

Speech emotion recognition is a typical classification task. Some predefined speech emotion feature sets are extracted during speech analysis step. In this work the extracted parameters are: prosodies, unvoiced frame, jitter, shimmer and MFCCs parameters to know the influence of each parameter chosen on the emotions fear, anger and neutral, for exploiting in the emotions recognition system. During the classification step the extracted feature sets are classified trying to define the emotional class of the analyzed speech patterns. The extraction of the 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 reminder of this paper is as follows. Section 2 presented notions on the emotion and the parameters used. The third section shows the database that used in our work, the extraction and the analysis of the selected parameters. In the fourth section, the recognition of the three emotions fear, anger and neutral is made with the method of K-Nearest Neighbor (KNN). Finally, Section 5 gives concluding remarks.

2. Notions on the Emotion and the Parameters Used

We present here a short definition of the emotion and the parameters which we chose in order to characterize the target emotions of our work.

2.1. 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].

2.2. The Parameters Used

The parameters used in this work are:

2.2.1. Prosodic Parameters

Prosodic features can be further classified in broadly three feature categories: Pitch, Intensity and Duration. The patterns of the pitch (fundamental frequency), duration and energy are very useful in the acoustic representation of the prosody [2]. Recently differenced prosody features [3] have been explored for the task of emotion recognition which has shown better performance when compared to the conventional prosody features.

The fundamental frequency F_0 or pitch: Pitch is one of the important features, also referred to as fundamental frequency. 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. The contour of fundamental frequency is represented in figure 1. 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 [4-8]. The combined spectral and differenced prosody features are considered for the task of the emotion recognition [9].

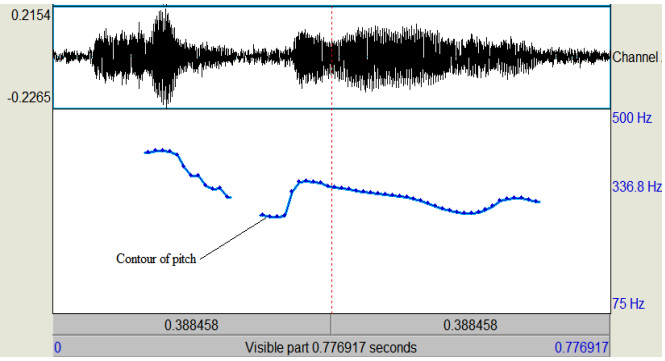

Fig. 1. Contours of fundamental frequency (pitch).

Intensity: Feature intensity refers to the energy or volume of the speech signal. Energy or volume depends on the loudness of the speech signal. 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 [10]. The figure 2 represents the contour of intensity.

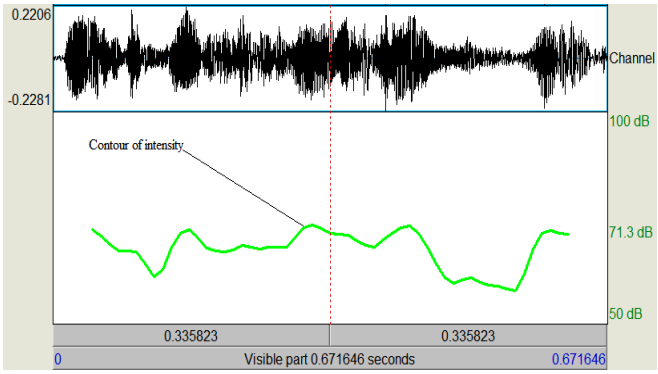

Fig. 2. Contours of intensity.

Duration: Duration is the most difficult parameter to specify because nothing indicates how the system of production control or speech perception measures time. The complex relations between pitch, duration and energy parameters are exploited for detecting the speech emotions [11].

2.2.2. Unvoiced Frames

The speech signal is broken up into a number of frames. These frames can be *voiced* or unvoiced. The voiced frames contain the prosodic parameters, and the unvoiced frames contain the parameters of excitation. Different wavelet decompositions are considered both for voiced and unvoiced segments, in order to determine a frequency band where the emotions are concentrated [12].

2.2.3. Jitter and Shimmer

Many acoustic measures of voice quality rely on quantifying the periodicity of a signal. Voice quality can be measured using jitter, shimmer and harmonic-to-noise ratio (HNR). Jitter and shimmer are perturbation measures, indexing the cycle-to-cycle variation in frequency and amplitude, respectively. A signal with higher jitter and shimmer is more variable and less periodic, representing poorer voice quality. While these measures are frequently reported in the literature, concern has been expressed regarding their relationship with perceived voice quality [13-14]. Different combinations of features: pitch, intensity, jitter, shimmer, spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) and formants relevant to emotions are explored [15].

2.2.4. Mel-Frequency Cepstral Coefficients

MFCCs parameters 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. These coefficients are derived from the cepstral coefficients and methods [16]. The MFCCs parameters are used in the speech emotion recognition [8], [17-18]. In [19], speaker emotions are recognized using the data extracted from the speaker voice signal. Mel Frequency Cepstral Coefficient (MFCC) technique is used to recognize emotion of a speaker from their voice. A novel feature fusion of Teager Energy Operator and Mel Frequency Cepstral Coefficients (MFCC), as Teager-MFCC (T-MFCC) feature extraction technique, is used to recognize the stressed emotions from speech signal [20]. To improve the performance, the study in [21] use the combination enter spectral and prosodic parameters. Information of fundamental frequency (F_0), the log of energy, the formants, energy in Mel and MFCCs are explored to classify the emotions. The combined MFCC and relative prosody features have yielded a 75% recognition rate [9].

3. Extraction and Analyzes of the Selected Parameters

In this section we present a definition for the database that used in our work, thus the extraction and the analysis of the selected parameters.

3.1. Database

To evaluate the performance of a speech recognition system it is essential to design a suitable database. Little work was done on Arabic language. An Arabic acted corpus composed of Tunisian dialect isolated words was created, in this study the first Arabic natural corpus of different dialects is proposed to recognize discrete emotions [22].

Algerian dialect is very different from Arabic dialects of the Middle-East, since it is highly influenced by the French language, , Berber and Turkish [23]. The database used is built from films in Algerian Dialect, this database contains 152 segments of duration ranging from 0.2 s to 3 s, and these segments include three emotions: fear, anger and neutral. Table 1 describes the number of segments for each emotion:

Table 1. Number of segments for each emotion

Emotions	**Fear**	**Anger**	**Neutral**
Segments number	52	52	48

3.2. Parameters extraction

Extraction of parameters is one of the important steps in the development of speech emotion recognition systems. The results obtained are given in table 2. This last illustrates the statistics values of the parameters chosen for the three treated emotions.

3.3. 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 table 2 that the statistical values mean and peak (max) of pitch are higher in fear compared to the anger and neutral states. The pitch range of anger emotion is the highest value.

In table 2 it shows for the intensity that the anger emotion has a highest mean value, the neural emotion has a broad range compared to the others states. It is observed that the duration concern to the emotion of fear is very short.

For analysis purpose speech signal is decomposed in to number of frames. In table 2 it is noticed that the emotion of anger has the lowest value number of unvoiced frames. But in the neutral and fear state the number of frames is moderate.

Table 2. The parameters extracted by software praat

Parameters	**Fear**	**Anger**	**Neural**
Mean of pitch(Hz)	**279.79**	269.09	190.60
Max of pitch(Hz)	**398.84**	387.36	273.25
Min of pitch(Hz)	163.18	131.87	**120.62**
Range of pitch(Hz)	235.65	**255.50**	152.62
Mean of Intensity(dB)	69.97	**71.29**	70.30
Max of Intensity(dB)	75.16	**77.35**	76.98
Min of Intensity(dB)	57.20	55.45	**51.11**
Range of intensity(dB)	17.96	21.91	**25.86**
Duration(s)	0.88	**1.34**	1.28
Unvoiced Frame (%)	24.12	23.92	**24.36**
Jitter (%)	**3.29**	3.19	2.89
Shimmer (%)	**16.03**	15.40	13.40

The values of jitter and shimmer are slightly higher in the fear emotion and lower in the neutral state.

MFCC is a very powerful technique to analyze the speech signal. In our work 20 Mel-Frequency Cepstral Coefficients are extracted. Figure 3 illustrates MFCCs forms of the three emotions fear, anger and neutral, these results it's obtained by Matlab software.

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

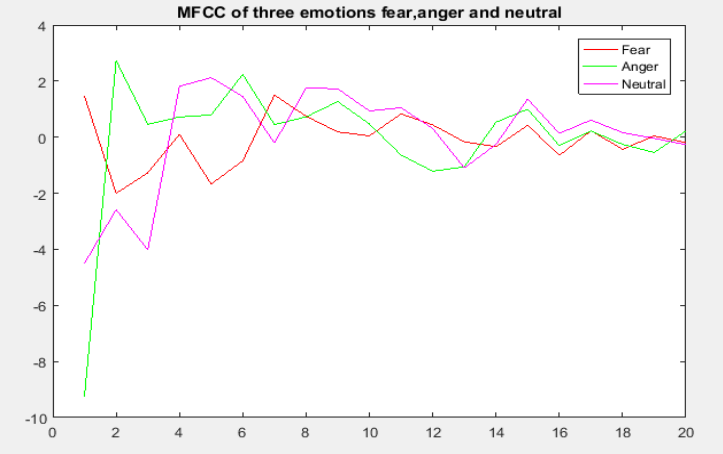

Fig. 3. MFCC of three emotions fear, anger and neutral.

4. Results and Evaluation

The emotions recognition systems are based on classifiers methods, of which we present a brief review on the method of classification KNN, we describe then the classification system of the emotional states related to the fear, anger and neutral.

4.1. Classifier

K-Nearest Neighbor (KNN) classifier is a type of instance based learning technique and predicts the class of a new test data based on the closest training examples in the feature space [24]. Euclidean distance was used as distance measurement [25]. KNN is simplest and influential method of classification [26]. Emotion recognition using speech processing using k-nearest neighbor algorithm [27]. In [28], emotions are recognized through speech using spectral and prosodic features, KNN and Gaussian mixture model (GMM) are used as classifiers.

4.2. The Emotions Recognition

The feature sets extracted in section 3 are used with KNN (k=3) based classifier for the evaluation of emotion recognition system. The system is divided into three parts. In the first part we use only the statistic values of prosodies parameters: the mean, the maximum, the minimum and the range of pitch, the mean, the maximum, the minimum and the range of intensity and the duration. The jitter, shimmer and unvoiced frames parameters are included in the second part. In the third part we insert the MFCCs parameters with the others parameters. We notice in table 3 the recognition rate of each part. Figure 4 shows a comparison between the recognition rates of each part.

Table 3. The performance of classification methods of each part.

Parameters	**Recognition rate**
Prosodies (pitch, intensity and duration)	**81.94%**
Prosodies + jitter + shimmer + unvoiced frames	**82.63%**
Prosodies + jitter + shimmer + unvoiced frames + MFCCs	**84.02%**

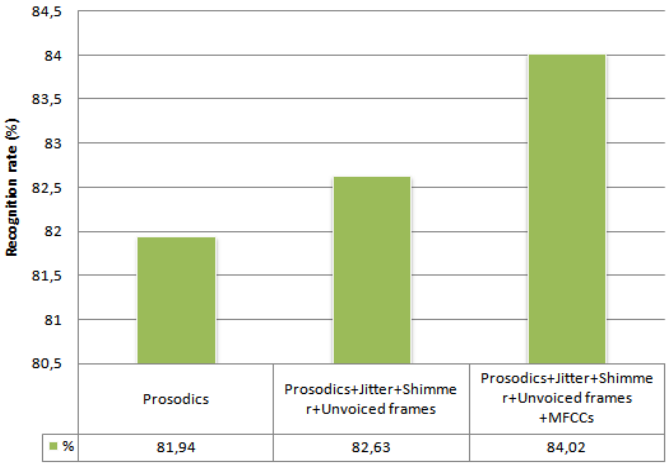

Fig. 4. Comparison between the recognition rates of each parts.

From the table 3 and figure 4 we notice that the obtained results are acceptable but vary according to the parameters that used. When we only used the prosodic parameters we find a recognition rate value around to 81.94%, there is an improvement of performance (82.63%) when the jitter, shimmer and unvoiced frames parameters are added in the recognition system. A higher recognition rate (84.02%) is obtained when the MFCCs parameters are included to the others parameters.

The confusions matrixes for each part were showed in tables 4, 5 and 6. According to the confusions matrixes, it is remarked that the fear and anger emotion have low recognition rates compared to neutral state. This remark can explain that the angry emotion and fear are almost similar in some of the parameters that we have previously mentioned in section 3, which correspond the mean and maximum of pitch.

Table 4. Confusion matrix of the first part (prosodies parameters)

Emotion	Fear	Anger	Neural
Fear	**79.17%**	12.50 %	8.33 %
Anger	8.33 %	**79.16%**	12.50 %
Neural	6.25%	6.25%	**87.50%**

Table 5. Confusion matrix of the second part (prosodies parameters+ jitter +shimmer +unvoiced frames)

Emotion	Fear	Anger	Neural
Fear	**81.25%**	10.42%	8.33 %
Anger	8.33 %	**77.09%**	14.58%
Neural	4.17%	6.25%	**89.58%**

Table 6. Confusion matrix of the second part (prosodies parameters+ jitter +shimmer +unvoiced frames +MFCC)

Emotion	Fear	Anger	Neural
Fear	**83.33%**	8.33 %	8.33 %
Anger	8.33 %	**77.09%**	14.58%
Neural	4.17%	4.17%	**91.66%**

5. Conclusion

This paper presents system of speech emotion recognition of three emotion fear, anger and neutral. For this purpose Algerian Dialect speech database is used for parameters extraction. We extracted the prosodic parameters, the unvoiced frames, jitter, shimmer and the MFCCs parameters concerning the three emotions treated. An analysis is made to know the influence of the three emotions on the extracted parameters. The system of recognition is based on the KNN method of classification. The obtained results show us that the used of the jitter, shimmer and unvoiced frames parameters with the prosodies parameters improve the performance of recognition (82.63%). Combined prosodies, jitter, shimmer, unvoiced frames and MFCCs parameters give classification rate important (84.02%). For future works, other emotion such sadness and surprise can be including in our emotion recognition system.

References

1. Marie Tahon. Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot, these doctorat, paris, 2012.
2. Rao KS, Yegnanarayana B(2006) . Prosody modification using instants of significant excitation. IEEE Trans Audio Speech Lang Process, 14(3):972–980.
3. Vidyadhara Raju VV, Gurugubelli K, Vuppala AK (2018). Differenced prosody features from normal and stressed regions foremotion recognition. In: 5th international conference on signal processing and integrated networks (SPIN), IEEE, 2018.
4. Schroder, M. . Emotional speech synthesis: A review. In Seventh European conference on speech communication and technology, Eurospeech Aalborg, Denmark, Sept. 2001.
5. Murray, I. R., & Arnott, J. L (1995). Implementation and testing of a system for producing emotion by rule in synthetic speech. Speech Communication, 16, 369–390
6. Wang, Y., Du, S., & Zhan, Y (2008). Adaptive and optimal classification of speech emotion recognition. In Fourth international conference on natural computation, (pp. 407–411).
7. 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.

8. K. Wang, N. An, B.N. Li, Y. Zhang, L. Li (2015), Speech Emotion Recognition Using Fourier Parameters, IEEE Transactions on Affective Computing, 6 , 69-75

9. Vegesna, V. V. R., Gurugubelli, K., & Vuppala, A. K. Application of Emotion Recognition and Modification for Emotional Telugu Speech Recognition. Mobile Networks and Applications, 2018.

10. Cowie, R., & Cornelius, R. R (2003). Describing the emotional states that are expressed in speech. Speech Communication, 40, 5–32.

11. Iida, A., Campbell, N., Higuchi, F., & Yasumura, M (2003). A corpus based speech synthesis system with emotion. Speech Communication, 40, 161–187.

12. Vasquez-Correa, J. C., Garcia, N., Orozco-Arroyave, J. R., Arias-Londono, J. D., Vargas-Bonilla, J. F., & Noth, EEmotion recognition from speech under environmental noise conditions using wavelet decomposition. International Carnahan Conference on Security Technology (ICCST), 2015.

13. Kreiman, J., Gerratt, B. R., & Gabelman, B (2002). Jitter, shimmer, and noise in pathological voice quality perception. The Journal of the Acoustical Society of America, 112(5), 2446.

14. Martin, D., Fitch, J., & Wolfe, V(1995). Pathologic voice type and the acoustic prediction of severity. Journal of Speech and Hearing Research, 38, 765–771.

15. Koolagudi, S. G., Murthy, Y. V. S., & Bhaskar, S. P (2018). Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. International Journal of Speech Technology, 21(1), 167–183

16. Bhoomika Panda, Debananda Padhi, Kshamamayee Dash,"Use of SVM Classifier & MFCC in Speech Emotion Recognition System", IJARCSSE ,Vol. 2, Issue 3,March 2012.

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

18. Bitouk, D., Verma, R., & Nenkova, A (2010). Class-level spectral features for emotion recognition. Speech Communication, 52(7–8), 613–625.

19. Likitha, M. S., Gupta, S. R. R., Hasitha, K., & Raju, A. U. Speech based human emotion recognition using MFCC. International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017.

20. Bandela, S. R., & Kumar, T. K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017.

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

22. Meddeb, M., & Alimi, A. Building and analyzing emotion corpus of the arabic speech. International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE, 2017.

23. Mohamed Menacer, Odile Mella, Dominique Fohr, Denis Jouvet, David Langlois, et al.. Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. ACLing 2017 - 3rd International Conference on Arabic Computational Linguistics, Dubai, United Arab Emirates. Nov 2017, pp.1-8.

24. R. O. Duda, P. E. Hart and D. G. Stork, Pattern classification. John Wiley and Sons. 2012.

25. M. Hariharan, Sazali Yaacob, M. N. Hasrul and Oung Qi Wei "speech emotion recognition using stationary wavelet transform and timbral texture features", ARPN Journal of Engineering and Applied Sciences vol. 9, no. 8, august 2014

26. Y. Song, J. Huang, D. Zhou, H. Zha, C. L. Giles (2007). IKNN: Informative K-Nearest Neighbor Pattern Classification, in: European Conference on Principles of Data Mining and Knowledge Discovery, 248–264.

27. Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta (2014). Emotion recognition using Speech Processing Using k-nearest neighbor algorithm. International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622.

28. Chandra Prakash, Prof. V.B Gaikwad, Dr. Ravish R. Singh Dr. Om Prakash (2015). Analysis of Emotion Recognition System through Speech Signal Using KNN & GMM. Classifier IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), 55-61.