EMOTION RECOGNITION FROM PERSIAN SPEECH WITH NEURAL NETWORK

Mina Hamidi[1]and Muharram Mansoorizade[2]

[1]Department of Computer Engineering, Bu-Ali-Sina University, Hamedan, Iran
m.hamidi@basu.ac.ir
[2]Department of Computer Engineering, Bu-Ali-Sina University, Hamedan, Iran
mansoorm@basu.ac.ir

ABSTRACT

In this paper, we report an effort towards automatic recognition of emotional states from continuous Persian speech. Due to the unavailability of appropriate database in the Persian language for emotion recognition, at first, we built a database of emotional speech in Persian. This database consists of 2400 wave clips modulated with anger, disgust, fear, sadness, happiness and normal emotions. Then we extract prosodic features, including features related to the pitch, intensity and global characteristics of the speech signal. Finally, we applied neural networks for automatic recognition of emotion. The resulting average accuracy was about 78%.

KEYWORDS

Emotion recognition, speech processing, emotional speech, recognition of anger, Persian language, Persian speech database

1. INTRODUCTION

Humans can communicate with each other through natural language. Machine should also be able to use the same way interact with humans, but still machines, can't understand all aspects and meanings of natural language. Moreover, identify and extract information from the human emotions (acts, facial expressions), for a real emotional connection without any discomfort is essential for human.

Both research and business are developing in the field of human interface technology which covers speech recognition, speech synthesis and virtual reality. However, there are many problems in affective information processing and recognizing human emotion accurately. This makes many people still feel strong resistance toward interacting with machines in business fields of terminal devices (mobile phones and car navigation systems) and medical care systems [4].

In recent years, personal robot technology and use it as a new way of training or just entertainment, great progress has been aimed. These robots are often like family dogs and cats are like Sony AIBO and sometimes like a small kid like Sony humanoids SDR3_X. Interact with this machine is different from old computers and must learn to communicate using new media.

DOI : 10.5121/ijaia.2012.3509

Robot must try to learn normal social conventions such as politeness and natural language. Among the capabilities these personal robots need, the most basic is the ability to grasp human emotions, and in particular, they should be able to recognize human emotions as well as to express their own emotions. Indeed, emotions are not only crucial to human reasoning, but they are central to social regulation and in particular to control dialog flows [8].

According to recent researches an increasing attention to speech signals has been found, so many systems have been proposed for detecting emotion from speech. Speech signal is the fastest and most natural method of communication between human, this fact has prompted researchers to speech as the fastest and most effective way of communication between humans and machines to think. Speech is the distinction between humans and other organisms and plays a major role in the expression and human communication. Moreover, in certain conditions such as communication via the phone, speech is only available channel [3]. Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where there sponse of those systems to the user depends on the detected emotion. It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety. It can be also employed as a diagnostic tool for therapists. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech. Speech emotion recognition has also been used in call center applications and mobile communication. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice [6]. But this requires that the machine is smart enough to recognize the human voice.

In recent seventy years, much research has been done on speech recognition, the human speech processing and converting it into a sequence of words referring to.

Although a lot of processing on speech recognition performance, but we are still far from having a natural interaction between human and machine, the machine does not understand human emotion states. This new research field has introduced a sense of speech recognition. Researchers believe that this sense of the speech recognition can be useful to extract meaning from speech and can improve the performance of speech recognition systems [6].

2. RELATED WORKS

Emotional recognition is a common instinct for human beings, which has been studied by researchers from different disciplines for more than 70 years. Fairbanks et al pioneering work on emotional speech revealed the importance of vocal cues in the expression of emotion, and the powerful effects of vocal emotion expression on interpersonal interaction. Understanding the emotional state of the speaker during communication can help the listeners to catch more information than is represented by the content of the dialogue sentences, especially to detect the 'real' meaning of the speech hidden between words. The practical value of emotion recognition from speech is suggested by the rapidly growing number of areas to which it is being applied, such as humanoid robots, the car industry, calling centers, etc. Although the techniques of machine learning and data mining functions are obtained under the boom, only a few works have used this powerful tool to better results in the sense of the letter reached [4].

According to Pantic and Rothkrantz 2003, the auditory features that are most often extracted from the speech signal are: (i) pitch, (ii) intensity, (iii) speech rate, (iv) pitch contour, and (v) phonetic features. Pitch corresponds to the rate at which vocal cords vibrate and determines the frequency of the acoustic signal, while intensity refers to vocal energy. Variations in voice pitch and intensity usually have a linguistic function, such as over-stressing or under-stressing certain words. When, for example, a person is experiencing anger, fear or joy, the sympathetic nervous system becomes aroused, resulting in a heart rate and blood pressure increase that produces mouth dryness and occasional muscle tremors. Speech is then characterised by loudness, increased speech rate and strong, high frequency energy. Speech rate represents the number of spoken words within a time interval. Finally, pitch contour corresponds to pitch variations described in terms of geometric patterns, and phonetic features of all types of sounds involved in a speech (e.g., vowels, consonants and their pronunciation) [2].

These systems achieved an accuracy rate of 72–85% when detecting one or more basic emotions from noise-free audiovisual input. These accuracy rates outperform the equivalent human emotion recognition skill that achieves an accuracy rate of 55–70% in neutral content speech [2].

3. PERSIAN EMOTIONAL SPEECH DATA BASE

Work in the field of speech recognition can be categorize in many ways. One of the problems in research on emotion recognition from speech there is a lack of appropriate database. In some languages, like English, there are more databases than other languages, but in between them, many databases are not available and can be personalized. Moreover, in most languages, including Persian, the database does not exist or is very limited.

Existing databases are classified according to different cases. For example, based on its language, the speech of children or adults, normal or abnormal speech, or number of emotion in data base. Due to differences in language structure and language conventions, or cultural differences among people that speaking different languages, speech evaluation and emotion recognition from speech is somewhat different from language to other language. To further explain these differences are expressed in the following examples:

Word in Persian language is not starting with silent but in English can.
In Persian language some consonant can't be together and all words have vowel phonemes, but other languages have no such restrictions

In other languages like English stress is part of the word but not in the Persian
Fumble in the Persian language, which expresses the speaker tries to hide something and he is confident but in a speech in some languages like French, this way of talking is part of normal speech.

According to the above matters and review the work has been done, feeling of need for research on emotion recognition from Persian speech, is quite palpable.

For this study, an appropriate Persian emotional speech database is needed, but unfortunately such a database was not available. Existing database or have made to speech recognition, such as speech database processing by the Sameti in Sharif University or are prepared to artificially, the sentences give to a specified person and they want to express such a way that they feel is needed in such an entirely recognizable like that Mansoori zade has done in his doctoral thesis, These databases represent abnormal emotional and sometimes the system has better result on this database than the real world and on the natural speech. According to the above mentioned reasons, we attempted to build Persian emotional speech database. The purpose of this database, use it for whatever you feel closer to the human natural speech and automatically recognition emotion from speech. This database made from actor or actress speech from more than 60 different films, because in film express much more natural emotion. Since, the number of different people speech who used in speech database is important[11], attempt to use more than 330 actor and actress. This database contains both men and women speech with different ages and contains over 2400 pieces of emotional Persian speech, including 6 feel anger, disgust, fear, sadness, happiness and normal mode. These utterances recorded with Pratt and length of them are different from each other.

4. EMOTIONAL RECOGNITION

The first step in recognizing emotion from speech is database. The next step is extraction appropriate features from speech.

These features are very extensive range of convenient and efficient features. We here use features consist of energy, rate, pitch and the MFCC coefficients. These features are extracted using the software Praat with great precision. If the speech signal as a time series x (t) is displayed where t is the time of expression, speech energy is defined as the unit of db is shown

E(t) = 10 log x(t)2 (1)

Intensity diagram of speech in normal and in a sense of anger has shown in Figure1.

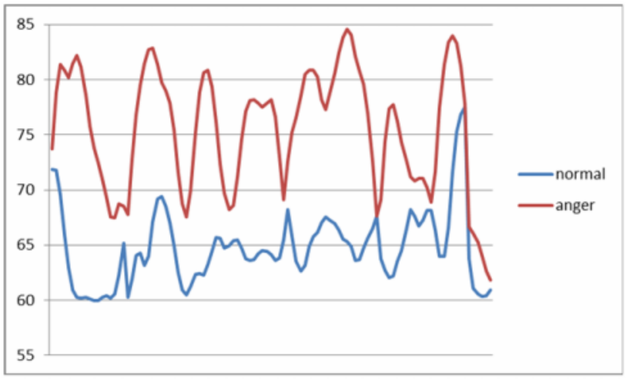

Figure1. Intensity diagram of speech in normal and in a sense of anger

Sounds that are produced naturally in the larynx are intermittent mode, in which period called frequency of the sound. Several methods, like correlation and frequency domain methods for detecting the frequency at each point of the speech signal, are exist and Praat has been implemented one of the most accurate algorithm[1].

Energy band that divided into frequency bands, especially on the mel scale have a special role in the show emotion in speech. MFCC is Fourier cosine transfer of signal amplitude and usually 10 to 12 members of its initial energy is used.

During the above processes of the original speech signal, energy, pitch, MFCC coefficients, its first derivative and also the minimum and maximum points and the normalized data that is calculated using the speaker's normal speech, are extracted.

The third step in the emotion recognition of speech is using the appropriate method to classify emotion. We use a multi-layer perceptron neural network. This network is composed of an input layer, a middle layer and an output layer. Network input is feature vector and its output is the class label. Network trained with input and binary label. During classification, the unknown samples given to the network, samples are awarded to the class that contains the maximum value. Result of each emotion has shown in figure2.

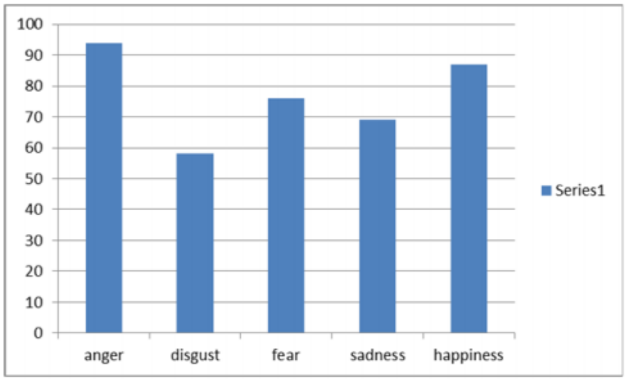

Figure2. Result for each emotion

Based on the above, average of accuracy is about 78%.

5. CONCLUSIONS

Since work on emotion recognition from speech recognition compared to other fields is a new feeling, so there are still numerous issues for study and research.

One of the considered issues is appropriate and useful database for each language that until have not or less database like Persian.

Database environment is another issue, noisy or without noise environment and natural or simulated speech in speech database.

Another case is, the number and types of emotions are in database. In real human life exist wide variety of human emotions, which can have different levels of intensity and weakness, what number of this emotion was more and has a wider range of changes, the work obtained more efficiently and more accurate result.

Another area that needs more study is the number and type of feature is used in research.
Size of the part of speech is very effective, speech can be either discrete or continuous and can has different length.

The first aim of this article is to build a proper database of emotional Persian speech that effort to be natural and with high quality.

In the second stage, simple and accurate features extracted to avoid the implementation complexity methods. Finally, with implementation and use of neural network reached to average accuracy about 78% , that in this field and with the conditions stated in this article is acceptable and appropriate.

REFERENCES

[1] گفتار و حرکات چهره, رساله یها یژگیبر و یص احساس انسان, مبتنیزاده, محرم, تشخ یمنصور
ت مدرس، 1388یوتر، دانشگاه تربیکامپ یمهندس یدکتر

[2] Lopatovska, I., & Arapakis, I. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. Information Processing and Management (2010), doi:10.1016/ j.ipm.2010.09.001

[3] Batliner,A. , Steidl,S., Schuller,B., Seppi,D., Vogt,T., Wagner, J., Devillers,L., Vidrascu,L., Aharonson,V., Kessous,L., Amir,N., Whodunnit , Searching for the most important feature types signalling emotion-related user states in speech , Computer Speech and Language 25 (2011) 4–28, Elsevier, doi:10.1016/j.csl.2009.12.003

[4] Rong,J., Li,G., Phoebe Chen,Y., Acoustic feature selection for automatic emotion recognition from speech, Information Processing and Management ,45 (2009) 315–328 ,Elsevier, doi:10.1016/j.ipm.2008.09.003

[5] Ren,F., Affective Information Processing and Recognizing Human Emotion, Electronic Notes in Theoretical Computer Science 225 (2009) 39–50, Elsevier, doi:10.1016/j.entcs.2008.12.065

[6] Ververidis,D., Kotropoulos,C., Emotional speech recognition: Resources, features, and methods, Speech Communication 48 (2006) 1162–1181, doi:10.1016/j. specom. 2006.04.003

[7] ElAyadi,M. , Kamel,M. , Karray,F., Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (2011) 572–587, Elsevier, doi:10.1016/j.patcog.2010.09.020

[8] Pierre-Yves,O., The production and recognition of emotions in speech: features and algorithms, Int. J. Human-Computer Studies 59 (2003) 157–183, Elsevier, doi:10.1016/S1071-5819(02)00141-6

[9] Yeh, J., Pao,T. , Lin,C., Tsai,Y., Chen,Y., Segment-based emotion recognition from continuous Mandarin Chinese speech, Computers in Human Behavior (2010), Elsevier, doi:10.1016/j.chb.2010.10.027

[10] Tawari,A., Trivedi,M., Speech Emotion Analysis in Noisy Real-World Environment, 2010 International Conference on Pattern Recognition,IEEE, DOI 10.1109/ICPR. 2010.1132

[11] Polzehl,T.,Schmitt,A., Metze,F., Wagner,M., Anger recognition in speech using acoustic and linguistic cues, Speech Communication(2011), Elsevier, doi:10.1016/j. specom.2011.05.002

[12] Chen,L., L. S. H. Chen, Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction, in Faculty of Graduate Studies Urbana Champaign: University of Illinis, 2000.

[13] De Silva, L. C. , Pei Chi, N. Bimodal emotion recognition, 2000, pp. 332-335.

[14] Busso, C. ,Deng, Z. ,Yildirim, S. ,Bulut, M. ,Lee, C. M. ,Kazemzadeh, A. , Lee, S. ,Neumann, U. ,Narayanan, S. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, in ICMI'04: ACM, 2004.

[15] Cheng-Yao, C. ,Yue-Kai, H. ,Cook, P. ,Visual/Acoustic Emotion Recognition, 2005, pp.1468-1471.

[16] Zeng, Z. ,Zhang, Z. ,Pianfetti, B. ,Tu, , J. ,Huang, , T. S. , Audio-visual affect recognition in activation-evaluation space, 2005, p. 4 pp.

[17] Paleari, M. ,Lisetti,C. L , Toward Multimodal Fusion of Affective Cues, in HCM'06: ACM,2006.

[18] Schuller,B. ,Arsic,D. ,Rigoll,G. ,Wimmer,M. ,Radig,B. , Audiovisual Behavior Modeling by Combined Feature Spaces, in proc. of ICASSP 2007, pp. II-733-II-736.

[19] Kessous,L. ,Castellano,G. ,Caridakis,G. , Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, Journal on Multimodal User Interfaces, vol. In Press, 2009.

[20] http://www.sharif.ir

[21] Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monograph, 6,87–104.

[22] Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. Speech communication special issue on speech and emotion (Vol. 40, pp. 5–32). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B.V.