International Journal of Digital Content Management
Allameh Tabataba'i University
Vol. 2, No. 3, Summer & Autumn 2021

Research paper

Persian Speech Emotion Recognition Approach based on Multilayer Perceptron

Seyed Mehdi Hoseini*

Date Received: 2021/08/25 Date Accepted: 2021/11/11

Abstract
Emotion recognition from speech has noticeable applications within the speech-processing systems. The goal of this paper is to permit a totally natural interaction among human and system. In this paper, an attempt is made to design and implement a system to determine and detect emotions of anger and happiness in the Persian speech signals. Research on recognizing some emotions has been done in most languages, but due to the difficulty of creating a speech database, so far little research has been done to identify emotions in Persian speech. In this article, because of the dearth of a suitable database in Persian to detect feelings, before everything, a database for moods of happiness and anger and neutral (with no emotion) in Persian, including 720 sentences was set up. Then the frequency features of speech signals obtained from Fourier transform such as maximum, minimum, median and mean as well as LPC coefficients were extracted. Then, the MLP neural network was used to detect emotions of happiness and anger. Results show that our algorithm performs 87.74% accurately.

Keywords: Emotion Recognition, Speech Processing, LPC Coefficients, Neural Network.

* MSc of computer science, Department of computer science, University of Mazandaran, Babolsar, Iran. Email: mehdihoseini.cs@gmail.com

Introduction

A speech identifies different emotional states of the person, including anger, happiness, surprise, fear, and so on. In this study, a distinction is made between the state of happiness, anger, and the state of neutrality (without any emotion), and a method has been proposed that, whenever a speech signal is given to identify it, it recognizes that it is said with a state of surprise or neutral.

Emotion recognition in speech is an important aspect of computer-speech communication (speech and speaker recognition, word combinations), which has been the reason for conducting experiments to develop effective algorithms for emotional state recognition in recent years.

So far, for many languages of the world such as English, French, Spanish, German, Italian, Dutch, Swedish, Danish, Arabic, Norwegian, etc., emotion recognition systems of speech signal with appropriate quality have been provided. However, for the Persian language, no significant work has been done other than recognizing anger and sadness [4], for other emotional states. This point shows the weakness in research and lack of attention of people to this language and research in the field of emotion recognition in Persian speech signal is necessary.

Garvian and Ahmadi (2008) extracted the speech parameters of Formant and step frequency for emotional states of anger and sadness in Persian and the speech mode was determined by decision tree and GMM methods.

Mosavian, Norasteh and Rahati (2008) studied feelings of hatred, anger, fear, sadness, happiness, and neutrality in Persian. To detect emotions, they used various acoustic features of pitch, jitter, shimmer, sound intensity and finally Fractal sound signal.

Mosavian, Norasteh and Rahati (2013) examined the effect of culture and social norms on expressing feelings of anger, happiness, sadness and neutrality when collecting data and recognizing emotions by considering the features of local culture in Persian. They used LPC length and coefficients along with length and frequency features and used the combined ANFIS method to detect emotions.

Arias, Busso and Yoma (2014) reviewed all speech recognition systems, between 2000 and 2013, with extraction features and all different databases and researches (in German, Chinese, Mandarin, Hindi, French, Slovenian and Spanish) (Hamidi & Mansoorizade,

2012), And then they define and categorize all the linguistic, non-linguistic, linear and nonlinear features and all the emotional states of hatred, anger, fear, sadness, happiness, surprise, fatigue, interest, anxiety, hostility, pride, satisfaction, hope and neutrality, and the various methods of analyzing different features and emotion recognition that have been proposed so far are described and compared.

According to the research, it is observed that some emotions such as anger, fear and sadness have been recognized for the Persian language, but the difficulty of creating a speech database and the lack of resources have made it impossible to detect the emotion of surprise for Persian speech. Therefore, with the aim of disseminating information in this field, the aim of this study is to find an efficient way to detect emotions of happiness and anger by examining LPC coefficients and signal frequency features.

This article consists of five sections. An overview of the proposed system is shown in Figure 1. The second part describes how to create and collect a database in Persian. In the third section, how to extract features and select the appropriate features will be presented. In the fourth section, the features are classified and the results obtained in this article will be presented and analyzed. In the fifth section, conclusions and future work will be presented.

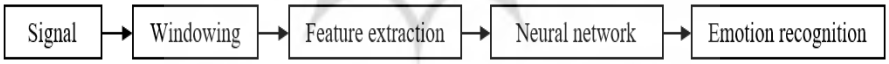

Figure 1: Overview of the proposed system

Materials and Methods

Databases

Speech emotion recognition systems require appropriate databases, standards, and frameworks in general. One of the problems in the field of research in the field of emotional processing of Persian speech is the lack or limitation of the emotional databases in Persian. There is no standard and well-known database for Persian, like other languages (Anagnostopoulos and et al., 2015 ; Bijankhan and et al., 1994).

Existing databases are classified according to different items, such as their language, children's and adults' speech, natural or simulated speech, or the number of emotions in the database. Due to differences

in the structure and rules of language and cultural differences among people, the evaluation and recognition of emotions from speech is somewhat different from one language to language. To further illustrate the differences, the following examples are given:

1. The word in Persian, unlike English, can not start with silence.
2. In Persian, some consonants cannot be together and all words have vowels, but other languages, such as English, do not have such a restriction.
3. In other languages, such as English, stress is part of the word, but not in Persian.

To prepare the database, The following actions were taken (Hamidi & Mansoorizade, 2012):

1. Design of Persian speech signal data
2. Creating a database in both emotional states of surprise and neutral

Design of Persian Speech Signal Data

One of the most important topics in this study is the collection of emotional data in Persian. There are different methods for collecting emotional speech data, the most important of which are:

Induction data: In this type of data collection, a scenario is usually predetermined and a conversation or sentence is considered for it, then a professional actor is asked to utter the relevant sentence according to the scenario and emotional state related to that sentence. This is the most common and simplest way to collect emotional data (Staroniewicz, 2011).

Real data: In this method, the behavior and expression of one or more speakers in a specific period of time is usually examined and recorded, then according to the different emotional states that the speakers have expressed during the study, the data is selected. In this way, the data is more real and the emotions are purer. Gathering data with this method is very difficult and time consuming and natural emotions are gently expressed. Sometimes it is difficult to recognize these emotions clearly. For this reason, this method is less used (Staroniewicz, 2009).

Video recording: This method uses emotional conversations in movies and TV and theater productions. This method is also less used. The problem with this method is the scattering of different sentences,

speakers, and emotional states that make it difficult to collect data and teach recognition systems (Ayadi and et al., 2011).

To design the data, a set of sentences prepared by the Intelligent Signs and Data Research Institute called "FarsDat" has been used. 15 sentences were randomly selected (Bijankhan and et al., 1994).

Database Preparation

With the help of 10 theater actors (5 women and 5 men), the selected sentences were expressed in three modes, with emotions of happiness, anger and without any emotion (neutral). Each actor was asked to put himself in the position of the scenario according to the desired emotion and the sentence related to it and the scenario that was predetermined for each sentence, and to express the desired sentence with the requested emotion.

Then each person uttered 15 sentences twice with a emotion of happiness, twice with a emotion of anger and twice in a neutral way in the laboratory environment and was recorded using Praat software in mono mode, 16 bits with a frequency of 44.1 kHz. The length of each speech is 22 seconds. Attempts were made to collect the best voices, with a total of 900 sentences recorded in three modes: happy, angry, and neutral. Silence between words in the said sentences was removed using Praat software and each sentence was saved in separate files (Parvinnia & Pourvahid, 2017).

Qualitative Evaluation of the Database

To ensure the high reliability and naturalness of the recorded sounds for the database, the listener perception test is performed, which after preliminary validation (evaluated by 4 recorded sounds), some suspicious sounds (which were not in the desired emotional state), or low-quality acoustic recorded sounds were removed, resulting in a happy-anger-neutral emotional database with 720 recorded records.

Speech Signal Preprocessing

Before processing the speech signal, it is necessary to make changes to it, including framing and windowing. The features of the speech signal and the speech duct change during the expression of a dialect; Therefore, the speech signal is a non-static signal and its statistical properties change over time. But because the organs of speech change state slowly, or in other words, man can not change them faster than a

certain limit, in small intervals, it can be assumed to be a signal of the station. For this reason, the speech signal is divided into short time intervals (usually 20 to 40 ms) and the signal analysis is performed on the signal at these short intervals. These parts of speech are called frames. On the other hand, in order to consider the boundary information of these sections, successive frames are selected as overlaps and the overlap rate is between 30% to 50% (Anagnostopoulos and et al., 2015). All the above steps can be done with the help of various toolboxes and commands of MATLAB software.

In this study, 256 sample frames with a length of 32 milliseconds and 50% overlap were selected. The sampling frequency is also 8 kHz. The window should be used to eliminate the effect of the edges. There are two advantages to using a window:

1. By weakening the signal at the beginning and end of the window, it reduces the effect of sudden change of amplitude at the beginning and end of the frame.
2. By multiplying the window in a speech signal in time, we will create a convolution of the window spectrum and the speech signal in the frequency axis, and this will eliminate the distortion resulting from the framing of the speech signal.

After segmentation, each window is multiplied by a window function before spectral analysis to reduce sample fragmentation. There are several types of windows, the most common of which are Hamming, Hanning, Rectangular and Blackman. Each window has advantages and disadvantages. The Hanning window was used in this study. Because the time signal of this window decreases on both sides, the sub-lobes in its spectrum are very weak and cause there is no strong ripple speech in the spectrum after this operation (Garvian & Ahmadi, 2008).

Extract Features

First, the effect of emotion on speech parameters should be obtained, and in the next step, appropriate parameters should be used to improve the recognition results. There is a very wide range of suitable and efficient features such as energy, speed, step, MFCC coefficients (Arias and et al., 2014) and fractal dimension (Mosavian and et al., 2013).

In this study, LPC coefficients and frequency features of speech signals such as maximum, minimum, middle and average have been considered to detect emotion. In the proposed method, first, each of the speech signals was normalized to extract linear prediction coefficients between zero and one, and then for each frame, 10 linear prediction coefficients were extracted using MATLAB software. From the coefficients obtained in each frame with other frames in each speech signal, mean, median, standard deviation, minimum, maximum were taken (Arias and et al., 2014), which for each linear prediction coefficient, 5 values are obtained and a total of 50 extraction features To be applied to the neural network. Then, for each speech frame, the first, second, third and fourth resonant frequencies were extracted. From the resonant frequencies obtained in each frame with the other frames in each speech signal, mean, median, standard deviation, minimum, maximum were taken. For each resonant frequency, 5 values are obtained and a total of 20 properties were obtained to be used as input to the neural network.

An appropriate method for classifying emotions should be used to identify the emotions expressed. The accuracy of the method used to identify speech mode depends on several factors. It is natural that the method used should be very accurate in identifying the desired condition. On the other hand, the ability of the method used to distinguish between two speech modes is also important. For example, one method may be very accurate in recognizing sadness, but it may not be very accurate in recognizing anger, and many speech frames may be identified as sad in anger. In this case, the labels that identify the speech window as sad can not be trusted in all cases. In fact, the best method is a method that has good accuracy for all cases (Lopatovska & Arapakis, 2011).

Multilayer Perceptron Neural Network

The multilayer perceptron neural network method, which is one of the most powerful techniques for classifying information, was selected in this study. The neural network used is a multilayer perceptron network. The important point in training multi-layer perceptron neural networks is to select the appropriate training algorithm that can be used to optimize the weights obtained using evolutionary algorithms (Addeh & Maghsoudi, 2016 ; Golilarz & Demirel, 2017).

Also, the number of hidden layers used is 2 hidden layers and the number of neurons used in the first hidden layer is 30 neurons and in the second layer is 20 neurons, which is obtained by trial and error by starting from a smaller number and The number of neurons has increased until the desired result is achieved. The extracted feature vector was given as input to the neural network, and its output, which is a emotion of happiness or anger or neutrality, was obtained. Network trained with inputs. In the classification of samples given to the network, the samples are assigned to the class containing the highest value.

Results and Discussion

To evaluate the proposed network, all the speeches in the database (both happy, angry and neutral) were given to the neural network and the result was compared with the input. MATLAB software divides the data entered into the network during the training process into two parts: training, testing and evaluation, and performs the training process in the training section separately.

The model is then evaluated using data that was not used in the training phase (test and evaluation set). It should be noted that increasing the number of speech sets in the database can play an important role in improving the results because the number of neural network training samples used increases.

The results presented in Table 1, which are obtained from an average of 10 repetitions of training, indicate that the proposed algorithm is suitable for both training and test sets. At a glance at its contents, it can be seen that the proposed method provides far better results than other methods used so far for other languages.

Table 1. Results of the Proposed Algorithm for the Created Database

The name of the set of speech / signals	Number of training data	Number of test data	Correct response rate
Emotion of Anger	119	36	89.03%
Emotion of Happiness	127	33	85.62%
Neutral (No Emotion)	74	31	88.57%
Proposed System	320	100	87.74%

To evaluate the performance of the proposed architecture, it is compared with the emotion recognition algorithms presented by Staroniewicz (2011) and Hamidi and Mansoorizade (2012).

Staroniewicz (2011) used the database of Polish emotional speech during tests. The features based on F0, intensity, formants and LPC coefficients were applied in seven chosen classifiers. Best results obtained for SVM. Results show that this algorithm performs 61.43% accurately.

Hamidi and Mansoorizade (2012) made an effort towards automatic recognition of emotional states from continuous Persian speech. they extracted prosodic features, including features related to the pitch, intensity and global characteristics of the speech signal. Finally, they applied neural networks for automatic recognition of emotion. The resulting average accuracy was about 78%.
Experimental results have shown that the proposed architecture performs better than competing algorithms and provides more accurate emotion recognition.

Conclusions

In this article, we have tried to provide a suitable solution for recognizing the emotions of happiness, anger and neutrality (without emotion) in the Persian speech signal. LPC coefficients along with maximum, minimum, median, mean and average frequency range are extracted from the created database set and using neural network has caused the recognition of emotion with high accuracy. Finally, the proposed system for the created database achieved an accuracy of 87.74%. Therefore, the proposed method provides better results than other methods used so far for other languages.

For future research, emotion can be distinguished from noise-free speech signals for better results. In addition to extraction features, other sound features such as Formant and step can also be used to detect emotion.

Recommended Citation

Hoseini, S. M., (2021). Persian Speech Emotion Recognition Approach based on Multilayer Perceptron. *International Journal of Digital Content Mangement*, 2 (3), 177-188.

References

Garvian, D., & Ahadi, S. M. (2008). Recognition of emotional speech and identification of speech in Persian. *ModaresTechnical and Engineering Journal*. Electrical Engineering Special Issue. 34. [in Persian]

Mosavian, E., Norasteh, R. & Rahati, S. (2013). Recognition of human emotions using neural-fuzzy network. *Proceedings of the 8th Intelligent Systems Conference.* Ferdowsi University of Mashhad, Mashhad, Iran. [in Persian]

Mosavian, E., Norasteh, R., & Rahati, S. (2008). Recognition of emotions in Persian speech using fractal dimension. *Proceedings of the 17th Iranian Conference on Electrical Engineering*. Iran University of Science and Technology, Iran, 8, 342-348. [in Persian]

Ranganath, R., Jurafsky, D. & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*. 27, 89-115.

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language*. 28, 278-294.

Hamidi, M., & Mansoorizade, M. (2012). Emotion recognition from Persian speech with neural network. *International Journal of Artificial Intelligence & Applications.* 3, 107-112.

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*. 43, 155-177.

Bijankhan, M., Sheikhzadegan, J., & Roohani, M. R. (1994). The speech database of Farsi spoken language. *Proceedings of the Australian Conference on Speech Science and Technology*, 2, 826-830.

Staroniewicz, P. (2011). Automatic recognition of emotional state in Polish speech. *In Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces*. Theoretical and Practical Issues, Springer, Berlin, Heidelberg, 347-353.

Staroniewicz, P. (2009). Recognition of emotional state in Polish speech-comparison between human and automatic efficiency. *In European Workshop on Biometrics and Identity Management*, Springer, Berlin, Heidelberg, 33-40.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*. 44, 572-587.

Parvinnia, E., & Pourvahid, M. (2017). Feature extraction from speech signals to identify the feeling Persian. *Proceedings of the National*

Conference on Computer Engineering and Information Technology. Islamic Azad University, Sepidan Branch, Iran. [in Persian]

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*. 47, 575-592.

Addeh, A., & Maghsoudi, B. M. (2016). Control chart patterns detection using COA based trained MLP neural network and shape features. *Computational Research Progress in Applied Science & Engineering*. 2, 5-8.

Golilarz, N. A., & Demirel, H. (2017). Thresholding neural network (TNN) based noise reduction with a new improved thresholding function. *Computational Research Progress in Applied Science & Engineering*. 3.