Designing the Intelligent System Detecting a Sense of Wonder in English Speech Signal Using Fuzzy-Nervous Inference-Adaptive system (ANFIS)

Sakine Tashakori
University of Sistan and Baluchestan
Sistan and Baluchestan Province, Zahedan, Daneshgah Boulevard, Iran
Phone: +98 54 3113 2505
sakine.tashakori@yahoo.com

Salman Haghighat
Andisheh University
Fars Province, Jahrom, Andishe St, Iran
salman.haghighat110@gmail.com

Abstract

The purpose of this research is to design an intelligent diagnostic system for detecting a sense of wonder in English speech signal using Fuzzy-Nervous Inference-Adaptive system (ANFIS). For English, the recognition of some surprise feelings such as anger, grief, joy and hatred has been made, but due to the difficulty of creating a speech database in a state of wonder and a shortage of resources in this case, even in other languages, so far, no sense of wonder has been detected in the English speech. In the absence of a suitable database in English for the identification of emotions, at first, a wonder-neutral database (without feeling) was created in Persian, containing 30 sentences with a sense of surprise and neutrality. Then, LPC coefficients and frequency characteristics of speech signals such as maximum, minimum, middle and mean (obtained by FFT) were extracted. Finally, the neuro-fuzzy adaptive network (ANFIS) was used to create a sense of wonder with an average accuracy of about 94.93%.

Keywords: Emotion Detection; English Speech Signal; Fuzzy-Nervous Inference-Adaptive System (ANFIS).

1. Introduction

One of the important features of speech is the transfer of the emotional state of the speaker to the listener. A speech describes various emotional states of an individual, including anger, happiness, surprise, fear, and so on. Understanding the sense of speech suggests more information to the listener in addition to its lexical meaning. Therefore, the listener is not only concerned with what the speaker says, but also the feeling along with it.

By increasing the transaction between man and machine, the need for automatic conversation between these two and the removal of the human operator is of particular importance. There is a lot of research to make it easier to communicate between them. Understanding human emotions from the machine and providing a good response to it is one of the areas that helps to reach this goal.

Creating a system for recognizing the sense of speech signal, due to the increased transaction between man and machine, plays an important role in everyday life and needs to increase more and more. In this regard, research is required to diagnose the sense of wonder in the English speech signal.

In Ghaderian and Ahadi's article (2008), the ferment speech parameters and step frequency for emotional states of anger and grief are extracted in Persian language and are determined by decision tree method and GMM speech mode.

Mousavian, Nourest, and Rahati (2007), investigated the influence of culture and social norms on anger, happiness, sadness and neutrality during data collection and recognition of emotions considering the characteristics of local culture in Persian language. The length and

coefficients of LPC are used along with the length and frequency characteristics and used to recognize the ANFIS combination method.

Nourest et al (2009) investigated the feelings of hatred, anger, fear, sadness, happiness and neutral state in Persian language. To recognize the sensation, various sound features under the waveform, jitter, shimmer, sound intensity and, finally, the fractal dimension of the sound signal have been used.

In Anagnostopoulos et al's paper (2012), all speech recognition systems worked between 2000 and 2011, with extractive features, and all databases and researches (in German, Chinese, Mandarin, Hindi, French, Slavic and Spanish). Then, all the linguistic, non-verbal, linear and non-linear features are defined and classified and all emotional states of hatred, anger, fear, sadness, happiness, surprise, fatigue, interest, anxiety, hostility, vanity, satisfaction, hope and humor, and a variety of methods for analyzing various characteristics and recognizing the feeling that has been suggested so far has been described and compared by Shashidhar and his colleges (2011).

Hamidi and Mansoorizad (2012) recognized feelings of anger, hatred, fear, sadness, joy for Persian language. Significance of speech signal, such as step and intensity, energy and MFCC coefficients, has been applied to detect feelings, MLP neural network with accuracy of 78%.

Pathak (2011) studied feelings of grief, anger, happiness, hatred, fear, and and used to obtain a better result in recognizing the sensation in the speech signal from neural networks.

Staroniewicz (2012 and 2009) studied six feelings of hatred, anger, fear, sadness, joy and surprise, along with neutral state in Polish. Different emotional features such as severity, formant and LPC coefficients have been analyzed with the help of neural network, supporting vector machine and decision tree.

Vogot et al (2008) studied various created databases in English and categorized various extracted features of speech, such as step, energy, and formant and finally, some methods for the analysis of various emotional features and sensory recognition such as neural network, supporting vector machine, decision tree and Markov secret model have been evaluated.

According to the research, it is observed that the recognition of some emotions, including anger, joy, grief for the Persian speech signal, has been made, but the difficulty of creating a speech database in a state of wonder has not caused the discovery of a sense of wonder for English speech yet. Therefore, with the aim of expanding the information in this field, the purpose of this research is to find an effective method for detecting a surprise by analyzing the LPC coefficients and signal frequency characteristics.

Yousefinejad et al. (2015) conducted a study titled "Detecting the sense of surprise" in the English speech signal. According to this research, it is challenging for computers to detect feelings. The main reason is the inability of the computer to understand the user's feelings. The purpose of this paper is to design a speech recognition system and provide a new method for improving the system. So far different features have been used in this regard, but none have practically linked the relationship between the range of sound and emotional states. Because the Bionic wavelet has more to do with this connection, it seems to be able to help separate emotional states. For this purpose, in this study, the bionic wavelet was used to extract the characteristic of audio signals in the automatic recognition of emotions from speech. The structure of the bionics is consistent with the structure of the human ear, and since human beings have a good understanding of speech sentiment, the use of a bionic wavelet can be useful for automatically detecting emotions from speech. The proposed structure has been evaluated on the Berlin database and Persian emotional speech data, which contains short sentences and expressions of negative emotions of fear, anger, discomfort, and normal state, and so on. The results of the experiments show that the proposed algorithm offers acceptable performance compared to the automatic recognition of emotions from existing speech.

Pourvhayed and Ayat (2012) conducted a study titled "Detection of sensation in Persian Speech Signal" using neural network. In this paper, attempts have been made to design and implement a neural network based system to determine and recognize the sensation of wonder in Persian singular speech. For Persian, some emotions such as anger, grief, joy and hatred have been

recognized, but due to the difficulty of creating a spoken word database in a state of wonder and a shortage of resources in this case, even in other languages, As a result, no detection of the sensation of wonder in the Persian language is done so far. Due to the inaccessibility of a suitable Persian database for emotion detection, a surprisingly neutral database (without feeling) was created in Persian, consisting of 260 sentences with a surprise and neutral sensation.

Marawi and the Ismailis (2014) conducted a research entitled Persian database for detecting feeling from speech. Today, one of the issues that play an important role in human-machine interaction is the recognition of sensation from the speech signal, so the use of a comprehensive database in the system for recognizing the feeling is important. So far, various databases have been provided in English, Danish, and other languages, but the Persian databases have not been seen so far. Therefore, in this article, Persian emoticons are presented as emotional drums for speech recognition. This database contains 748 sentences with 8 feelings of anger, fatigue, hatred, fear, natural discomfort, surprise and joy. The sentences are expressed by 33 speakers (18 men and 15 women). In order to evaluate and compare the proposed database and the famous Berlin database, various features from the sentences of these two databases were extracted.

Ebrahimpour and Mahmoudian (1394) conducted a research entitled Detection of speech emotions using feature selection based on recursive models. Today, recognizing the sense of speech in cases where there is a relationship between human and machine are considered. Despite many efforts in this field, there is still a long way between the natural feelings of humans and the perception of the computer. In this article, the database of Berlin data as the most famous database available with 550 sentences created by professional actors in the lab environment, of which 61 sentences have been used with different feelings such as happiness, hatred, neutral, fear, discomfort, anger and fatigue. Various features of the sentences of this base are extracted individually and due to the large number of features, a method is needed to reduce the space of the property before applying the classification algorithm. To this end, a back-up vector-based recursive method (SVM) has been developed to extract the effective features in recognizing the sensation of existing data. The median diagnostic value is obtained only with the use of eight attributes more effectively than among the 75 existing attributes.

This article contains five sections. The second part describes how to collect the database in English. The third part will show how to extract the appropriate features. In the fourth section, the results obtained in this paper will be presented and analyzed. In the fifth part, the conclusions and future work will be discussed.

2. Database

One of the problems that exist in the field of emotional speech processing in English is the lack of or limited emotional database in English. Unfortunately, there is no standard and known database in English language as the other languages (Anagnostopoulos, 2012; Shashidhar, 2012; Lugger and Yang, 2007). The following measures were taken to prepare the database:

- Designing English Speech Signals Data
- Preparation of databases in two emotional states: wonder and neutral

2.1. Designing the Persian Sign Language Signal Data

For the design of the data, a set of sentences prepared by the researcher has been extracted and used among professors and learners. 12 sentences were chosen randomly.

2.2. Preparation of databases in two emotional states: wonder and neutral

With the help of 10 professors and EFL learners (5 professors and 5 EFL learners), the selected sentences were expressed in two emotional states, with a sense of wonder and without any feelings (neutral).

Each person expressed twelve sentences twice, and tried to collect the best sounds, altogether 240 sentences were captured in two states: wonder and neutral.

Then the silence between the words in the sentences was deleted, using the PRAAT software, and each of the sentences was stored in separate files.

2.3. Quality assessment of the database

To ensure the high reliability and the naturalness of the sounds recorded for the database, the listener's test is performed, which after the preliminary validation (evaluated by 3 recorded voices), some suspicious sounds (not in the emotional state) or low-quality acoustic sounds were omitted and eventually a wonder-neutral database with 200 recordings was obtained.

3. Feature extraction

There is a very wide range of suitable and efficient features such as energy, velocity, step, MFCC coefficients (Shashidhar, 2012) and fractal dimension (Noorest et al, 2009).

In this study, LPC coefficients and frequency characteristics of speech signals such as maximum, minimum, middle, and mean are used to detect sensation.

In the presented method, each of the signal sentences in the database is initially divided between zero and one, and then each sentence is divided into several subclasses. For each sub word, a suitable number of LPC coefficients were calculated using MATLAB software. Then, by applying the Fourier transform on subclasses, frequency characteristics such as maximum, minimum, middle, mean, mode, medium of frequency range, and domain for each one were extracted. In total, 47 features were used to recognize speech.

4. Detection of feeling

To identify emotions, an appropriate method for classifying emotions should be used. If adaptive neural fuzzy network is selected, the selection of a suitable adaptive neural network (ANFIS) and appropriate instructional training has effect in the improvement of system performance. Multi-layer adaptive neural network fuzzy network method, one of the most powerful techniques for classification of information, was selected in this research.

Several parameters such as the number of hidden layers, the number of neurons in each hidden layer, and educational algorithm are effective for teaching multilayer perceptron neural network. After examination, the neural network of the multilayer perceptron with two layers of cover was considered and 47 features extracted from the speech signal were given as input, in its first layer there were 24, and in the second layer there were 12 neurons. An emotional database was created from 200 existing sentences, 160 sentences for teaching and 40 other sentences for network testing.

Since the dispersion of data is very high, we first normalize it to minimize the error rate.

Initial implementation:

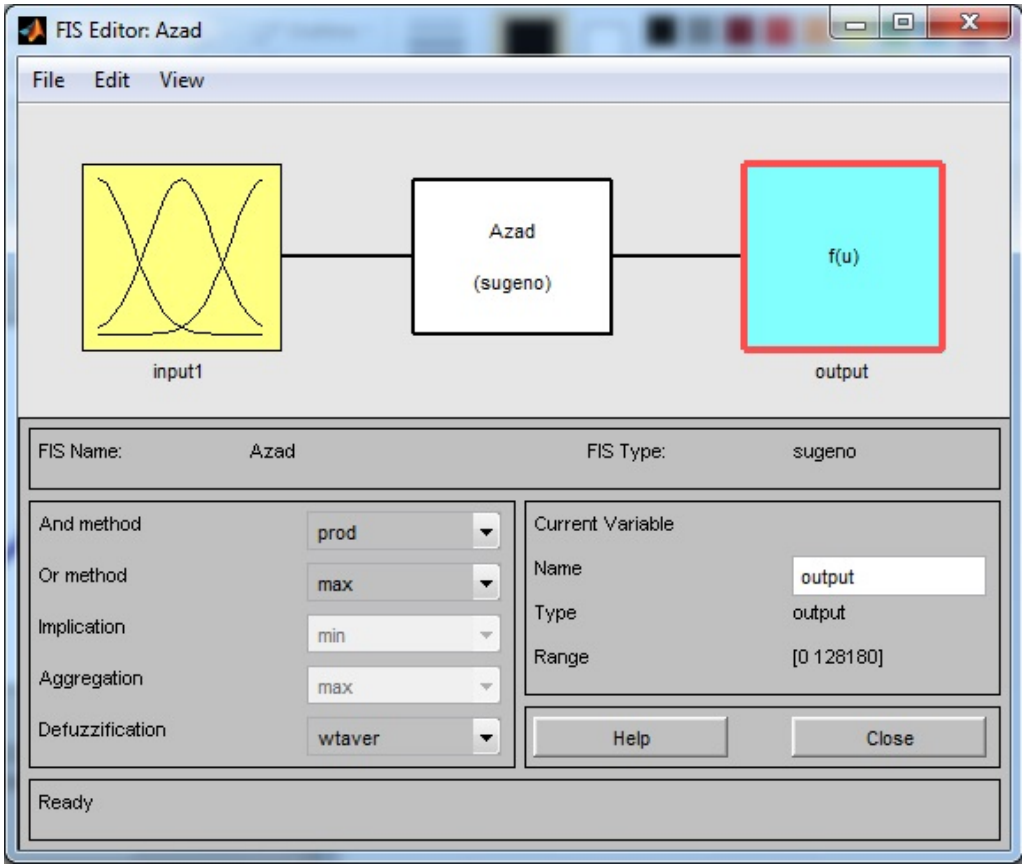

Figure 1. Initial implementation of the application

Created Model Structure: The structure of the created neural network has an input, The input, depending on its nature, has a number of different membership functions, which directs them to an output, and eventually 12 rules are derived.

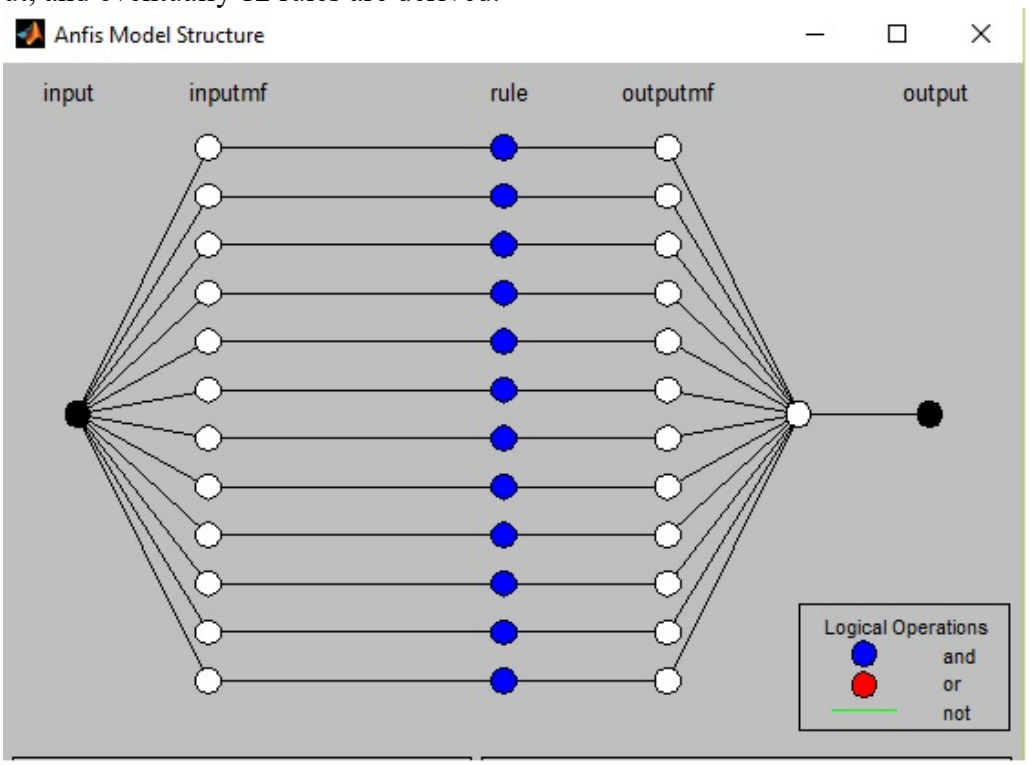

Figure 2. Network structure created for system training

In the provided ANFIS model, the error rate reaches zero according to the input data after 10. This confidence level is obtained by the ANFIS model, which provides 70% of the input data as the training to the network and keeps the remaining data for testing the suggested model. The training of the network proceeds to a point where the error reaches zero or a small amount that is inclined to zero. In our proposed network after the epoch, the error rate reached an ideal value of zero, which we could conclude 12 fuzzy rules.

Finally, the overall conclusion is that with more features, more accurate estimation of the degree of emotion detection can be achieved.

According to the results, 12 rules are obtained:

- If (input1 is in1mf1) then (output is out1mf1) (1)
- If (input1 is in1mf2) then (output is out1mf2) (1)
- If (input1 is in1mf3) then (output is out1mf3) (1)
- If (input1 is in1mf4) then (output is out1mf4) (1)
- If (input1 is in1mf5) then (output is out1mf5) (1)
- If (input1 is in1mf6) then (output is out1mf6) (1)
- If (input1 is in1mf7) then (output is out1mf7) (1)
- If (input1 is in1mf8) then (output is out1mf8) (1)
- If (input1 is in1mf9) then (output is out1mf9) (1)
- If (input1 is in1mf10) then (output is out1mf10) (1)
- If (input1 is in1mf11) then (output is out1mf11) (1)
- If (input1 is in1mf12) then (output is out1mf12) (1)

In order to quantify the precise amount of orders, there is a need to phase out a set of fuzzy rules that are trained with the help of ANFIS. Designed with the help of a GUI, this is easily possible and is displayed by placing each number in the desired range in the output.

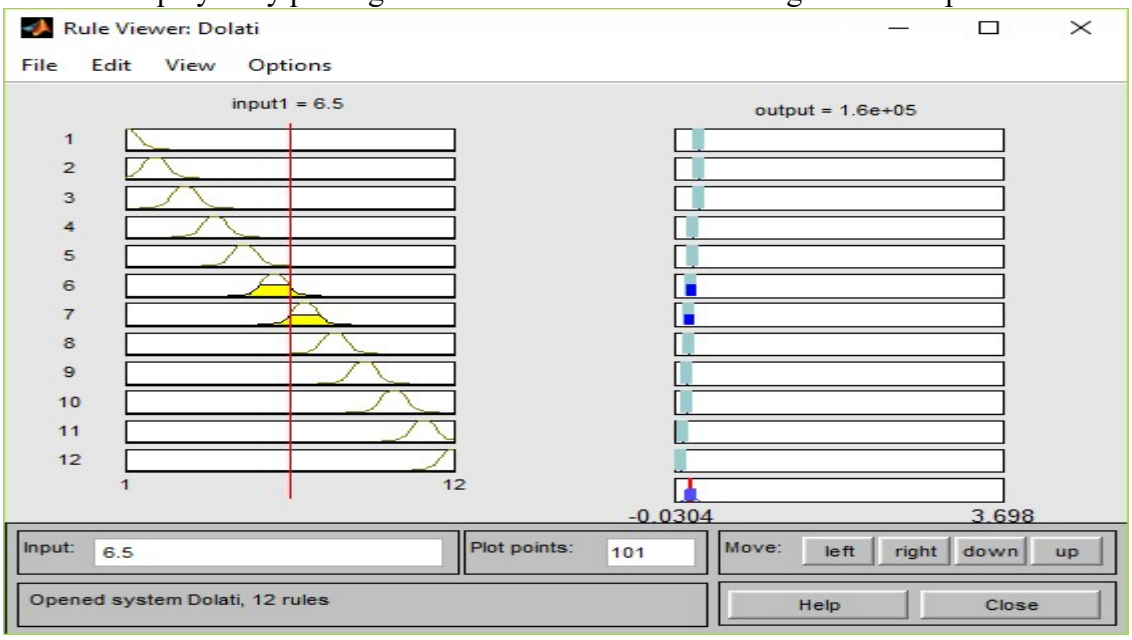

Figure 3. Shape of the Rules for Estimating

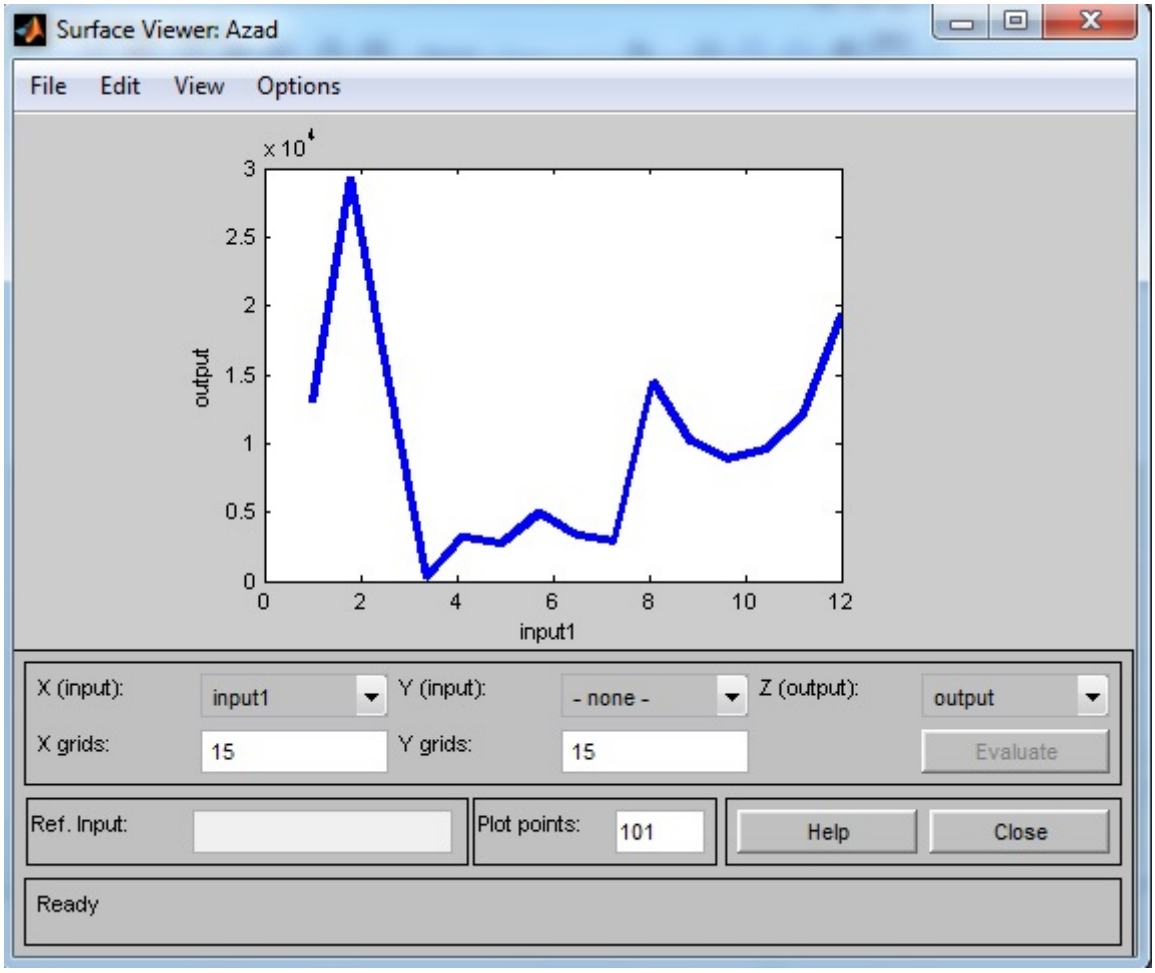

Figure 4. Trend Line Forecast, Detecting a sense of wonder In the English speech signal

In order to evaluate the network, all the speeches in the database (both wonder and neutral) were given to the ANFIS network and the result was compared with the input. 10 errors were identified from the instruction set to the system, and after giving the test set it counted 5 errors. Based on the above, the average accuracy of the system is about 94.23%.
The results presented in Table 1 indicate that the proposed algorithm is well-suited for both the training and the test set.

In Table 1, the results of the proposed method have been compared with other studies. With a glance at its contents, one can find that the proposed method offers far better results than other methods used so far for other languages. It should be noted that increasing the sample of the existing set of speeches in the database can play a significant role in improving the results, as the number of ANFIS network training samples is increased.

Table 1. The results of the proposed algorithm for the created database

The name of the set of speech signals	Number of errors in detection	Number of set	Correct answer percentage
Educational Speech of Emotional Signal Set	10	200	95.15%
The whole emotional speech of database	15	240	94.23%

Table 2. Comparing the results with similar research

Research done	Answer percentage
In (Hamidi et. al., 2012)	60 %
In (Nourest et. al., 2009)	57.7%
Suggested method	94.23%

Table 3. Compare the results with other learning machines

The name of the emotional speech database data base	Correct answer percentage
ANFIS	%94.23
Fuzzy	%84.13
ANN	%91.49
Deep learning	%96.63
(Reinforcement learning)	%95.86

The model of fuzzy systems is based on building a scientific basis using expert knowledge. However, in some processes where human experiences are not available, it is difficult to construct this scientific basis, so if the initial parameters of the comparative system are properly applied by the expert, the convergence rate of the parameters to their optimal values, as well as the convergence of the output response to the optimum route, significantly increases. One of the most suitable methods for setting parameters is the adaptation of neuro-fuzzy networks.

In the end, the results of the ANFIS model with the ANN model were compared and it was observed that the ANFIS model, due to the use of fuzzy rules, has more capability to ANN models in predicting the sensation of surprise in the English speech signal. The results of this research are consistent with the results of the above mentioned studies. They also concluded that the fuzzy-nerve inference method had a higher accuracy than the neural network method, But other machine learning techniques, such as deep learning, provides better results for training.

5. Conclusion

In this article, we have tried to find a suitable way to detect a sense of wonder in the English speech signal. The LPC coefficients are extracted with the maximum, minimum, middle, mean and average frequency range from the created database, and using the nerve fuzzy adaptive network has led to a wonder sensation with high accuracy and speed. Finally, the proposed system for the database reached to 94.23% accuracy.

For future research, you can detect a sense of wonder from the noiseless speech signal to get better results. Also, in addition to extraction features, other sound features such as Formant and step can also be used to detect a sense of wonder.

References

Anagnostopoulos, Ch.N., Iliou, Th., & Giannoukos, I. (2012). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Springer Science+ Business Media.*

Ebrahimpour, B., & Mahmoudian, H. (2014). Speech emotion detection using feature selection based on recursive models, 7th National Conference on Electrical and Electronic Engineering, Islamic Azad University, Gonabad, Iran.

Gharayan, D., & Ahadi, M. (2008). Recognition of emotional speech and speech mode identification in persian language. *Modares Technical and Engineering Magazine, Special Issue of Electrical Engineering*, No. 34 .

Hamidi, M., & Mansoorizad, M. (2012). Emotion recognition from Persian speech with neural network. *International Journal of Artificial Intelligence & Applications (IJAIA)*, *2 (5),* 107-112.

Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. *IEEE ICASSP. IV*, 17–20.

Marwi, H., & Ismailian, Z. (2014). Introduction of Persian databases to detect feeling of speech, Shahrood University of Technology, Iran.

Mousavian, E., Nourest, R., & Rahati, S. (2007). Recognition of human emotions using a neural network-fuzzy. Eighth Conference of Intelligent Systems, Ferdowsi University of Mashhad.

Nourest, R., Rahati, S., Sharifi, Sh., & Mousavian, E. (2009). Recognition of Sentiment in Persian Speech Using Fractal Subjects. Proceedings of 17th Iranian Conference on Electrical Engineering, Iranian University of Science and Technology.Vol. 8, 348-348.

Pathak, S. (2011). Recognizing emotions from speech. *Electronics Computer Technology (ICECT). 4*, 107 – 109.

Pouroohid, M., & Ayat, S. (2012). Detection of the sensation of surprise in persian speech signal using neural network, *Second National Computer Conference*, Sanandaj, Sama Faculty of Engineering

Shashidhar G., Koolagudi, K., & Sreenivasa, R. (2012). Emotion recognition from speech: a review. *Springer Science+Business Media, 15*, 99–117.

Staroniewicz, Piotr. (2009). Recognition of Emotional State in Polish Speech – Comparison between Human and Automatic Efficiency. *Springer, Heidelberg, 5707*, 33–40.

Staroniewicz, Piotr. (2011). Automatic recognition of emotional state in Polish. *Springer, Heidelberg,* 347–353.

Vogt, T., Andre, E., & Wagner J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization. *Springer, Heidelberg, 4868*, 75–91.

Yousefinejad, R., Haji Bagher Naeini, B., & Shafieian, M. (2015). Detection of the sensation of speech signal using the bionuclear wavelet, *Scientific Journal of Sound and Vibration, Vol. 5,* No. 9.