Designing and Implementing of Intelligent Emotional Speech Recognition with Wavelet and Neural Network

Bibi Zahra Mansouri
Department of Computer Engineering, Kerman Branch, Islamic Azad University
Kerman, Iran

Hamid Mirvaziri*
Department of Computer Engineering,University of ShahidBahonar
Kerman, Iran

Faramarz Sadeghi
Department of Computer Science, University of Shahid Bahonar,
Kerman, Iran

***Abstract*—Recognition of emotion from speech is a significant subject in man-machine fields. In this study, speech signal has analyzed in order to create a recognition system which is able to recognize human emotion and a new set of characteristic has proposed in time, frequency and time–frequency domain in order to increase the accuracy. After extracting features of Pitch, MFCC, Wavelet, ZCR and Energy, neural networks classify four emotions of EMO-DB and SAVEE databases. Combination of features for two emotions in EMO- DB database is 100%, for three emotions is 98.48% and for four emotions is 90% due to the variety of speech, existing more spoken words and distinguishing male and female which is better than the result of SAVEE database. In SAVEE database, accuracy is 97.83% for two emotions of happy and sad, 84.75% for three emotions of angry, normal and sad and 77.78% for four emotions of happy, angry, sad and normal**

Keywords—Recognition of emotion from speech; feature extraction; MFCC; Artificial neural network; Wavelet

I. INTRODUCTION

Speech is a communicative process among humans. One of the most significant characteristics of speech is transferring of internal emotion to the audiences. When the speech is presented by the speaker, the speech includes the individual's emotion. In this study, the researcher is going to recognize individual emotion. Recognizing emotions mean understanding speaker's emotion by speech's samples. It is better to use suitable parameters for improving he result of emotional speech. Firozshah et al have used MFCC and ANN to recognize four emotionsas angry, happy,neutral and sad which have the accuracy of recognition 72.05, 66.05 and 71.25 for women, men and mixtures of themrespectively [1]. Javidi et al have used MFCC, ZCR, Pitch, Energy and combination of the CHAID decision Tree, Regression, SVM, C5.0 and ANN to recognize as angry, happy, neutral, sadness, disgust, fear and boredom emotions, and the accuracy of recognition using ANN was 71.70 [2]. Dai et al [3], have presented neural network and combination of feature as a landmark, Pitch, energy to recognize speech's emotions as angry, happy, neutral and sadness and the accuracy of 90% was obtained for recognizing angry and neutral and more than 80% accuracy was obtained for sad and happy and more than 49% was obtained for classifying four emotions. Ayadi et al have worked by feature extraction ANN and HMM. The number of emotions was 7 for them. The accuracyrate was 71% for HMM and 50% for ANN which has shown better operation of HMM. [4]. Haq et al have used 7 emotions of angry, disgust, fear, happy, neutral, sad, surprised and energy feature extraction, duration, MFCC, pitch and MLB which has obtained 53%accuracy rate [5]. Ververidis et al have used angry, happy, neutral and sad emotions. They extract the features of energy, formant and pitch and their accuracy was 53.7%[6]. Gharavian et al have used GMM model and used four emotions and its accuracy was 65.1%. In this study, they used modular neural- SVM and applied happy, angry and neutral's emotions. The accuracy rate was 76.3%. In this study, the accuracy rate for C5.0 was 56.3% [7]. Table 1 shows the previus works in this field. This article organized as follows: In section two emotion speech reconition system is introduced. In section three feature extraction is stated. proposed method is in section four and it is evaluated with different dataset in section five and six and finally there is a conclusion and future works in the last two sections.

TABLE I. PREVIOUS MODELS AND THEIR RESULTS IN RECENT YEARS

Previous work	Emotion	Feature	Classifier	Database	Recognition rate
Ververidis D, Kotropoulos c (2006) [6]	Anger, happiness, neutral, sad, surprise	Pitch, Energy, Formant	MLB	DES-SUSAS	53.7%
Ayadi M, Kamel S, Karray F (2007) [4]	Anger, disgust, happiness, neutral, anxiety, tiredness	Energy, MFCC	ANN, HMM	EMO-DB	71% 55%
Dai, K, Fell, H.J, MacAuslan, J(2008)[3]	Hot-Anger, happiness, sadness, neutral, cold anger	Pitch, energy, landmark	ANN	Emotional Prosody Speech and Transcripts corpus	49%
Haq S, Jackson PJB, Edge J (2008) [5]	Happiness , anger, disgust, fear, sadness, surprise, neutral	Pitch, Energy, MFCC	MLB	EMO-DB	53%
Firoz Shah. A, RajiSukumar. A,BabuAnto. P(2010)[1]	Anger , happiness, neutral, sadness	DWT	ANN	INDIAN - DB	72,05% 66.05% 71.25%
Javidi M, Roshan E(2013)[2]	Anger , disgust, happiness, neutral, sadness, fear, bordem	Energy, Pitch, MFCC, ZCR,	SVM, ANN, C5.0,	EMO-DB	71.7%
Gharavian D, Sheikhan M, (2013)[7]	Happiness , anger, neutral, sad	MFCC, formant , pitch	GMM,C5.0, MLP, MODULAR NEURAL - SVM	PERSIAN - DB	65.1%56.3%76.3%

II. EMOTION SPEECH RECOGNITION SYSTEM

The emotion recognition system includes four main parts. "Fig. 1"shows the information of emotion recognition system.

The Pattern and emotion recognition system include four main processes, which are as the following: speech input, feature extraction, classifier, emotion speech output.

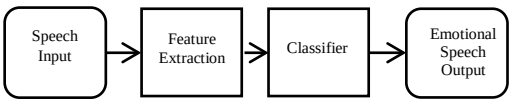

Fig. 1. Emotion speech recognition system

III. ANALYSIS AND FEATURE EXTRACTION

Extraction and selection of the best parameters of the speech signal are the most significant duty in designing a speech recognition system. Fig. 2 shows the preprocessing steps of speech analysis and feature extraction.

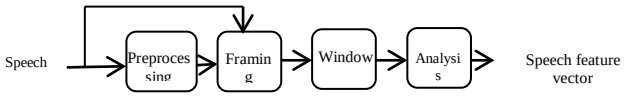

Fig. 2. Block diagram of speech signal analysis

A. Framing

When an audio vector is analyzed, the features are divided into two parts, half of them are in audio frame and the rest is in the frames. It is probable that the features are not achieved completely in each window analysis and they probably are hidden. Since, after converting analog signals to digital one, the speech samples are divided into frames in order to overlap each other. The new frame includes part of the previous frame and the next one [8].

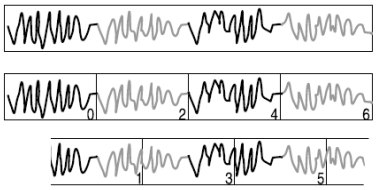

Fig. 3. Frame speech signal with overlapping frames

B. Windowing

Due to the non-static hood of speech signal and variable statistical features during time, the speech signal is divided into short times about 4-20ms and analyzing is conducted during mentioned time, these speech are called window.

After framing all frames at the beginning and end of each frame include interruption that is spectral distortion are reached to the least by framing at the beginning and end of each frame [9].

N is the number of symbols in each frame and n is the number of frames. Then the result of framing is the equation (1)

$$y(n) = x(n)\, w(n) \quad (1)$$

X(n) is the input signal of speech and y(n) is the output of the framing. Windows include varied models and some of them are introduced in the following equations [2].

$W(n) = \left[1 - cos\left(\frac{2\pi n}{N-1}\right)\right] 0 \le n \le N\text{-}1$ Hanning (2-a)

$W(n) = 0.54 - 0.46\ Cos\left(\frac{2\pi n}{N-1}\right)\ 0 \le n \le N\text{-}1$ Hamming (2-b)

$W(n) = 10 \le n \le N\text{-}1$ Rectangular(2-c)

C. Energy

Energy is the most and significant and basic features in speech signal which recognize the boundary between speech and silence. The energy can be obtained as follows [2].

The energy (E) of a signal frame of length N is obtained by

$$E = \sum_{n=0}^{N-1} y(n) * y(n) \quad (3)$$

D. Zero Crossing (ZCR)

The calculation of ZCR is done with audio signal which recognize the speech signals from silence [10].

The ZCR Crossing of a signal frame of length *N* is obtained by

$$ZCR = \frac{1}{N}\sum_{n=0}^{N-1} \frac{|sgn[y(n)] - sgn[y(n-1)]|}{2} \quad (4)$$

Sgn [y(n)] = 1 if y(n) ≥0
Sgn [y(n)] = -1 if y(n) < 0

E. Mel Frequency Cepstral Coeifficient (MFCC)

The main aim of using MFCC is inspiring from human's ears feature in receiving and understanding speech. The operation of human's ear is in a way that understanding frequency is varied from real frequency. One Mel is the measurement unit of heard frequency of a phoneme. It doesn't rely on pitchfrequency linearly, since the operation of human's ear is in a way that it doesn't understand more than 1 kHz frequency auditory system of human doesn't understand the frequencies ever scale linear, since the researcher has presented Mel scale for developing human understands. Mel frequency is a logarithmic mapping of physical frequency to understand frequencies. MFCC's coefficient considers certain coefficient for each frequency and since varied emotions considering different morality and mood have different frequency; therefore, anger is different from happiness and using these features increase the strength of emotion recognition.

F. Pitch

The periodic information of thin and thick speech is recognized mainly by pitch frequency. The more pitch the more thin sound and the less pitch frequency the more thick sound.

This frequency, which is called base frequency and it is shown by F_0 , is about 50 to 150 hertz in men. In women it isabout 150 to 450 hertz and in children it is about 300 to 700 hertz.

One of the oldest ways of estimating pitch in speech is autocorrelation. In this method autocorrelation changes of the function r (η) are plotted with respect to η (sample frame) [11].

$$r(\eta) = \sum_{n=0}^{N-1-\eta} y(n) * y(n-\eta) \quad (5)$$

G. Discrete wavelet transform(DWT)

The most usual signal analysis is Fourier transform which break up signals to different frequencies and keep the information of frequencies and lose the information of time, while the waveletincludes both of them that is, frequency information and time – oriented information. Equation of wavelet transform has presented in the following[1].

$$W_{j,k} = \sum_j \sum_k s(k).2^{-\frac{j}{2}} \Psi(2^{-j}n - k) \mathrm{k}, \mathrm{j} \in z \quad (6)$$

$\Psi(t)$is the main wavelet of analysis function and S(k) is speech signal, j is time measurement, k is the amount of movement in each measurement (transform parameter) and $w_{j,k}$ is wavelet coefficients.

In discrete wavelet transform, the main signal passes through the low-pass and high–pass filters which are appeared as approximation and detailed coefficients. In speech signals, low frequency is known as approximation h(n) , and high frequncey is known as details g(n) which has shown in Fig.4.

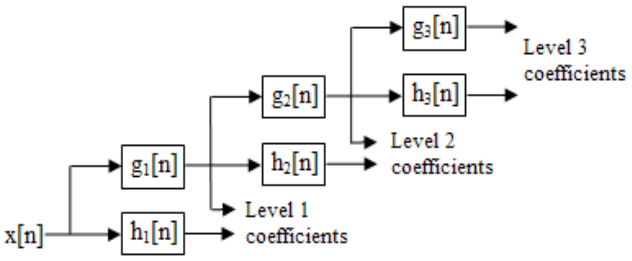

Fig. 4. Analysis of discrete wavelet on three level[12]

IV. Classifying Model of Artificial Neural Network

A multilayer neural network (MLP) shows a non-linear relationship between input and output vectors. It operates through connecting neurons of each node to the previous and next layers. The output of each neuron is multiplied by the weighting coefficients and is given as input to non-linear functions. In training phase, educational information is given to perceptron. Then weights are adjustedso that error between the output current and target reduces to minimum or the number of training reaches the predetermined one. Then to evaluate the accuracy of the training process, a series of inexperienced input is applied to the network.

MLP architecture consists of an input layer, hidden layer and output layer each of them includes specific neurons. The numbers of neurons of input layers is equal to the vector's features plus bias neuron and the number of output neurons is equal to the defined class for classifier. It ispossible to change the number of middle neurons till the best accuracy is obtained. In this research the number of middle neurons is changed from one to five and the best result has been obtained when the network has just had one hidden layer. Therefore a hidden layer of 12 neurons were considered.

V. Database

Emotional database in emotion recognition is usually applied for studying acoustic, phonetic and research and development in the field of emotion speech recognition system. SAVEE and EMO-DB have been studied for this research.

A. Berlin Database of Emotional Speech (EMO-DB)

This database is produced in technical university of Berlin. Seven emotional moods have used in this database which are angry, happy, boredom, sadness, fear, disgust and neutral. Ten artists who were five men and five women run this program. The number of speeches is about 535. The number of audio files for seven emotions is classified as follows [13]: angry (127), boredom (81), disgust (69), happy (71), sadness (62) and neutral (79).

B. Surrey Audio-Visual Expressed Emotion Database (SAVEE)

The database consists of the recorded voice of four made actor in 7 different emotions and 480 British speeches which has chosen amongst TIMIT database. These databases have recorded on forms of audio, video and audio-video. The database has recorded four British speakers of surrey's postgraduate student between the ages of 27 to 31. The samples are 44.1 KHz for audio and 60 fps for video. The number of audio files for seven emotions classified as follows [14]: angry (60), surprise (60), disgust (60), fear (60), happy (60), sadness (60) and neutral (120).

VI. Implementation Method and Analysis of Results

In this study data mining is done by ANN classifier and IBM SPSS software.

IBM SPSS Modeler is one of the best data mining tools and professional software to perform complex calculations and statistical analyzes for server and client.

Our data include 60 features extracted from a Berlin database speech and SAVEE. The outputs were angry, happy, sad and neutral. In this study, 340 speeches are chosen from Berlin database and 300 speeches are chosen from SAVEE. 20 percent of data were used for testing and 80 percent were used for training. This process has been repeated many times and the accuracy of classification is obtained based on the samples which recognized rightly to the all samples. Then, average of accuracy values,calculates for all repetition and presented as final accuracy (5 fold cross validation).

In the first experiment, recognizing of two emotions was performed. Our emotional moods were happy and sad. The whole number of speeches in Berlin database were 132 which include happy (70), sadness (62). The number of speeches was 120 in SAVEE database which include happy (60) and sadness (60). The result of each test for all features and features combination has done and their accuracy is presented in table 2. As you see in the table, the wavelet features in EMO-DB is 85.29% and in SAVEE is 53.57% by using the feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, pitch Fourier accuracy is obtained 100% in EMO-DB and 97.83% in SAVEE database, and for the Berlin database, accuracy is 2.7% better than SAVEE.

In table 2 the accuracy of emotions happy and sad in IBM SPSS Modeler Software in Berlin database obtained as 91.43%, and after removing lost data in MFCC features, it changes to 100%.

In the second test, it is tried for three emotions of angry, sad and neutral. All of 69 features were used in the test. The whole number of speech in Berlin database was 267; 126 for angry, 79 for neutral and 62 for sadness. The speech numbers were 240 in SAVEE database, 60 for angry, 120 for neutral and 60 for sadness. The result of each test has done for all features and features combination which their accuracy is shown in table 3. As you see in the table, for the wavelet feature is 63.64% in Berlin database and 54.24% in SAVEE database and by using feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, and pitch Fourier, it is obtained as 98.48% in Berlin database and 84.75 in SAVEE that is, in Berlin database 13.73% is better than SAVEE.

TABLE II. The Accuracy of Happy and Sad

Dataset	SAVEE	EMO-DB
WAVELET+FFTZCR	78.57	88.24
WAVLET+FFTPITCH	82.14	88.24
WAVELET+FFTENERGY	78.57	91.18
WAVELET+ZCR	92.86	97.06
WAVELET+ENERGY	89.29	100
WAVELET+PITCH	89.29	97.06
WAVELET+MFCC	67.86	58.82
FFT ZCR	85.71	88.24
FFT pitch	78.57	85.29
FFT energy	78.57	91.18
Energy	92.86	97.06
ZCR	96.43	97.06
Pitch	92.86	97.06
MFCC	71.43	61.76
Wavelet	53.57	85.29
Mixture of features	97.83	100

TABLE III. The Accuracy of Angry, Sad, Neutral

Dataset	SAVEE	EMO-DB
WAVELET+FFTZCR	62.71	78.79
WAVLET+FFTPITCH	72.88	71.21
WAVELET+FFTENERGY	74.58	74.24
WAVELET+ZCR	76.27	93.94
WAVELET+ENERGY	69.49	89.39
WAVELET+PITCH	72.88	87.88
WAVELET+MFCC	62.71	42.42
FFT ZCR	57.63	84.85
FFT pitch	79.66	65.15
FFT energy	74.58	81.82
Energy	77.97	81.82
ZCR	76.27	95.94
Pitch	86.44	84.85
MFCC	62.71	46.97
Wavelet	54.24	63.64
Mixture of features	84.75	98.48

In table 3, the accuracy of angry, sad and neutral emotions in IBM SPSS Modeler software in Berlin database firstly obtained 96.97%, and after removing lost data in MFCC feature it changes to 98.48%

In the third test, it is tried to classify angry, happy, sad, neutral emotions include 60 features. The whole number of speeches in Berlin database were 337 which include 126 angry, 79 neutral, 62 sadness and 70 happy. In SAVEE database, the speech number was 300 which include 60 angry, 120 neutral, 60 sadness and 60 happy. The result of eachtest for all features and feature combination has done and their accuracy has presented in table 4. As you see inthe table, for wavelet feature, it is 56.25% in Berlin database and 37.5% in SAVEE, and by using the feature combination as wavelet, MFCC, energy, ZCR, pitch, energy Fourier, ZCR Fourier, pitch Fourier, it is obtained as 90% in Berlin database and 77.78% in SAVEE and Berlin database accuracy is 12.22% more than SAVEE.

TABLE IV. THE ACCURACY OF ANGRY, HAPPY, SAD, NEUTRAL

Dataset	SAVEE	EMO-DB
WAVELET+FFTZCR	50	62.5
WAVLET+FFTPITCH	61.11	56.25
WAVELET+FFTENERGY	65.28	67.5
WAVELET+ZCR	61.11	86.25
WAVELET+ENERGY	56.94	65
WAVELET+PITCH	65.28	70
WAVELET+MFCC	43.06	41.25
FFT ZCR	55.56	62.5
FFT pitch	63.89	47.25
FFT energy	76.39	63.75
Energy	70.83	70
ZCR	63.89	88.75
Pitch	79.17	67.5
MFCC	47.22	34
Wavelet	37.5	56.25
Mixture of features	77.78	90

In table4, the accuracy of happy, sad, angry and neutral emotions in IBM SPSS Modeler software in Berlin database firstly obtained as 76.5%, and after removing lost data in MFCC feature changes to 90%.

VII. CONCLUSION

In this study, EMO-DB emotional speech is used which mad at the technical Berlin University and SAVEE emotional speech, made in Surrey University IN England. In this database, the data have high quality. The next step in recognizing emotional speech is feature extraction, that wavelet, MFCC, energy, pitches, ZCR. In this study, the feature combination of time – frequency, time, frequency domain and ANN classifier is used and feature combination in EMO-DB for two emotion is 100%, for three emotion is98.48%, and for four emotion is 90%, which all are better than SAVEE accuracy of recognizing emotional speech as 97.83% for happiness and sadness, 84.75%, for angry, sad and neutral and 77.78%, for happiness, sadness, angry, neutral.

VIII. SUGGESTION FOR FUTURE WORKS

It can be concluded from the mentioned studies that by a combination of features, the accuracy increases, which is significant, but it is not the best method since it is time consuming. Regarding this reason, the researcher has chosen the combination of neural network with Evolutionary Algorithm which improves the quality of speech emotion recognition.

REFERENCES

[1] F.Shah, A.R. Sukumar, andB.Anto, "Discrete wavelet transforms and artificial neural networks for speech emotion recognition," International Journal of Computer Theory and Engineering, 2(3), 2010, pp. 1793-8201.

[2] M.M .Javidi, and E.F. Roshan," Speech Emotion Recognition by Using Combinations of C5. 0, Neural Network (NN), and Support Vector Machines (SVM) Classification Methods," Journal of Mathematics and Computer Science, 6, 2013, pp. 191-200.

[3] K. Dai, H.J.Fell, andJ.MacAuslan, " Recognizing emotion in speech using neural networks,"Telehealth and Assistive Technologies, 2008, pp. 31-38.

[4] E.Ayadi, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in: The proceedings of the international conference on Acoustics, Speech, and Signal Processing, vol 5, 2007, pp. 957-960.

[5] S.Haq, P.J.Jackson, and J.Edge, "Audio-visual feature selection and reduction for emotion classification," In: The proceedings of international conference on Auditory-Visual Speech Processing, 2008, pp. 185-190.

[6] D.Ververidis, and C.Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," In: The proceedings of European signal processing conference,2006, pp. 1-5.

[7] M.Sheikhan, M.Bejani, and D.Gharavian, " Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method," Neural Computing and Applications, 23(1), 2013, pp. 215-227.

[8] I.McLoughlin, "Applied speech and audio processing: with Matlab examples," Cambridge University Press. 2009

[9] J.S.Devi, Y. Srinivas, and S.P. Nandyala, "Automatic Speech Emotion and Speaker Recognition Based on Hybrid GMM and FFBNN," International Journal, 2014.

[10] L.R. Rabiner, and R.W.Schafer, "Introduction to digital speech processing, Foundations and trends in signal processing," 1(1), 2007, pp. 1-194.

[11] X.Li, "SPEech Feature Toolbox (SPEFT) Design and Emotional Speech Feature Extraction," Faculty of Graduate School, Marquette University, 2007.

[12] S. Sunny, S.David Peter, andK.P. Jacob, "Performance Analysis of Different Wavelet Families in Recognizing Speech," International Journal of Engineering Trends and Technology, 4, 2013, pp. 512-517.

[13] .http://emodb.bilderbar.info/docu/#emodb

[14] http://kahlan.eps.surrey.ac.uk/savee/Evaluation.html