Automatic Emotion Recognition from Speech using Artificial Neural Networks with Gender-Dependent Databases

Firoz Shah.A, Raji Sukumar.A, Babu Anto.P
School of Information Science and Technology, Kannur University, Kerala, India
firozathoppil@gmail.com, rajivinod.a@gmail.com, bantop@gmail.com

Abstract: Automatic Emotion Recognition (AER) from speech is one of the most important sub domains in affective computing. We have created and analyzed two emotional speech databases from male and female speech. Instead of using the phonetic and prosodic features we have used the Discrete Wavelet Transform (DWT) technique for feature vector creation. Artificial neural network is used for pattern classification and recognition. We obtained a recognition accuracy of 72.055% in case of male speech database and 65.5% recognition in case of female speech database. Malayalam (one of the South Indian languages) was chosen for the experiment. We have recognized the four emotions neutral, happy, sad and anger by using Discrete Wavelet Transforms (DWT) and Artificial Neural Network (ANN) and the performance for the two databases are compared

Key words: Automatic Emotion Recognition, Affective Computing, Discrete Wavelet Transform, Artificial Neural Networks, Multi Layer Perceptron,

I. INTRODUCTION

Affective computing is concerned with emotions and machines. By giving machines the ability to recognize different emotions and make them able to behave in accordance with the emotional conditions of the user, human computer interfaces will become more efficient. Affective computing is a new field of modern advanced scientific research and is closely related to human emotional behavior and neural information processing. Emotions are well rooted to perception and the human neural system. Only through emotions, the communication will be effective. Emotions can be expressed and identified through speech, facial expressions, gestures etc. Since speech is the most effective medium for communication, speech emotion recognition attains greater importance [1]. Human speech is a combination of linguistics and emotions. Automatic emotion recognition from speech is a difficult pattern classification problem. Speech-based studies are context and database dependent. The emotional datasets can be speaker dependent and speaker independent. The speech corpus for emotion-based studies can be classified into natural, acted and elicited datasets. Making a machine that is able to respond with emotions is still a challenging goal. The most studied attributes for both speech and speaker based studies are MFCCs, LPCs, Fundamental frequency F0, formants, voice energy etc. If the machines can understand the emotional contents they can behave in a more friendly way. Emotions are the most important aspect in cognition [2, 3]. Automatic Emotional Recognition (AER) finds applications in speech recognition systems, text to speech synthesis systems, forensics, medical domains and humanoid robots. Emotions are closely related to psychological, linguistics, performance of the emotion recognition system depends on the features that are used. We have used Discrete Wavelet Transform (DWT) for feature extraction in this work.

II. A DISCRETE WAVELET TRANSFORM

In Discrete Wavelet Transform (DWT) a time-scale representation of a signal is obtained by digital filtering techniques. The DWT is computed by successive low-pass and high-pass filtering of the discrete time domain signal. Discrete Wavelet Transform uses filter banks to construct a multi resolution time-frequency plane. In DWT a discrete signal x[k] is filtered by using a high-pass and a low-pass filters, which will separate the signals to high and low frequency components, so that the output will contain half of the frequency contents, but the equal amount of samples as the original input signal[4]. To reduce the number of samples in the resultant output we apply a down sampling factor of ↓2. The DWT is defined by the following equation.

$$W\ (j,\ K) = \Sigma_j \Sigma_k X\ (k)\ 2^{-j/2} \Psi\ (2^{-j} n-k) \qquad (1)$$

where Ψ (t) is the basic analyzing function called the mother wavelet

At each level the decomposition of the input signal have two kinds of outputs which forms the low frequency components. i.e. approximations and high frequency components, the details [5]. The filtering and decimation process is continued until the desired level is reached the DWT of the original signal is then

978-0-7695-3915-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ACT.2009.49

obtained by concatenating all the coefficients, a[n] and d[n], starting from the last level of decomposition as shown in Figure1.

The successive filtering of the high pass and low pass filtering of the signal can be depicted by the following equations [6, 7]

$$Y_{high}[k] = \Sigma_n x[n]g[2k-n] \quad (2)$$

$$Y_{low}[k] = \Sigma_n x[n]h[2k-n] \quad (3)$$

where Y_{high} and Y_{low} are the outputs of the high pass and low pass filters obtained through sub sampling by 2.

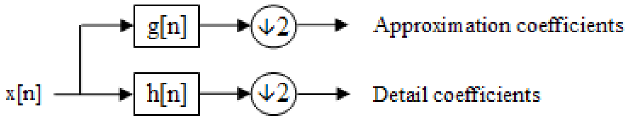

Fig 1 : Wavelet Decomposition

III. Classification And Recognition

Artificial Neural Network (ANN) is a system which operates in a very similar manner as that of the human brain, and solves problems by self learning. Artificial Neural Network is a computational model for information processing which combines artificial neuron in order to process information. By adjusting the weight of an artificial neuron we can obtain the output we want for a particular input. The process of adjusting the weights is known as learning [8]. Neural network can classify recognize convert and learn parameters. Multi Layer Perceptron is a supervised learning network hence it can use the information containing in patterns for efficient pattern classification. The MLP is a feed forward neural network consisting of nodes arranged in layers with only forward connections to units in subsequent layers. The connections have weights associated with them. Each signal traveling along a link is multiplied by its weight [9]. The input layer, being the first layer, has input units that distribute the inputs to units in subsequent layers. In the following (hidden) layer, each unit sums its inputs and adds a threshold to it and nonlinearly transforms the sum (called the net function) to produce the unit output (called the activation).The output layer units often have linear activations, so that, output activations is equivalent to function values.

IV. Dataset

Two elicited context databases were created for the experiment by using the Malayalam language. First database created consists of 340 male speech samples and the second database consists of 300 female speech samples.

V. Experiment And Results

We conducted two experiments to recognize the four different emotions neutral, happy, sad and anger from speech. Two elicited databases were created for the experiment by using male speakers and female speakers. Both the databases are speaker independent. We have used speakers under the age group of 30 for recording the speech corpus. A high quality studio recording microphone was used for the recording purpose. The speech samples are recorded at a frequency range of 8 KHz (4 KHz band limited). The speakers are trained well before recording the speech corpus. The recorded speech samples are processed labeled and stored in the dataset. For the feature extraction purpose we have used Daubechies-8 type wavelet. By using Daubechies-8 wavelet we performed the successive decomposition of the speech signals to obtain a good feature vector. The obtained feature vector is used to train MLP is recognizing four different emotions from speech. The database is divided into two for training and testing respectively. We used a proportion of 80% for training and remaining 20% for testing of the classifier in the case of both the databases. The male speech database we analyzed consists of 340 utterances and female dataset consists of 300 speech samples. We have used 10 male and 8 female speakers to create the databases.

A. *Experiment 1*

In the first experiment we used the male speech database. After successful training of the neural network by the feature vector from the 13th level of decomposition by using Db8 wavelet we have performed the testing of the network for recognizing the four different emotions; the recognition accuracies indicated in the confusion matrix is as shown in table1. While testing the network for recognizing the four different emotions neutral, happy, sad and anger the machine obtained a maximum recognition accuracy in the case of the emotions anger and least for happy. Whenever the machine tried to recognize the neutral speech from the four different emotional classes the machine obtained a recognition accuracy of 76.47%, while the machine faced a confusion of 17.64% with the emotion happy and a confusion of 5.88% as sad and the machine faced no more confusion with the emotion anger. For recognizing the emotion happy the machine can attain only a recognition accuracy of 52.94% and faced confusion of 17.64 % with the emotion neutral, 17.6% with the emotion sad and a confusion of 11.76% with the emotion anger. In recognizing the emotion sad the machine attained a recognition accuracy of 70.58% and faced a confusion of 17.64% with the emotion neutral, a confusion of 11.76% with the emotion sad and no more confusion with the emotion anger. For recognizing the emotion anger the machine attained a recognition accuracy of 88.23% and a confusion of 11.76% with the emotion neutral and there is no more confusion occurred in the case of emotions happy and sad. We could achieve an overall recognition accuracy of 72.055% from this experiment

Emotional Class	Neutral	Happy	Sad	Anger
Neutral	76.47%	17.64%	5.88%	0%
Happy	17.64%	52.94%	17.6%	11.76%
Sad	17.64%	11.76%	70.58%	0%
Anger	11.76%	0%	0%	88.23%

Table 1: Confusion Matrix obtained in experiment1

B. Experiment 2

In the second experiment we used the female speech database. After successful training of the neural network we have tested the neural network with the testing data. During testing the machine can recognize the emotion neutral with a recognition accuracy of 60% and faced an equal confusion of 13.3% with happy, sad and anger. While trying to recognize the emotion happy from the different emotional classes the machine can achieve a recognition accuracy of only 46% and confused with 20% with emotion neutral, 13.3% with sad and 20% confusion with the emotion anger. In recognizing the emotion sad the machine obtained a recognition accuracy of 60% and faced a confusion of 20% with neutral, a confusion of 13.3% with the emotion happy and obtained a confusion of 6.7 % with anger. While trying to recognize the emotion anger the machine obtained a recognition accuracy of 100% and the machine faced no more confusion with the emotion neutral, happy and sad. We can achieve an overall recognition accuracy of 66.5% from this experiment. A confusion matrix indicating the recognition accuracies for different emotions are given in Table2

Emotional class	Neutral	Happy	Sad	Anger
Neutral	60%	13.3%	13.3%	13.3%
Happy	20%	46%	13.3%	20%
Sad	20%	13.3%	60%	6.7%
Anger	0%	0%	0%	100%

Table 2 : Confusion Matrix obtained in experiment2

The recognition accuracy obtained for the four different emotions in both male and female speech database is shown in Fig (2). In the case of the emotions neutral, happy and sad, male speech obtained the maximum recognition. For recognizing the emotion angry the female speech attained the maximum recognition.

VI. CONCLUSION

We used the Discrete Wavelet Transform (DWT) as the feature extraction mechanism for automatic recognition of four different emotions .neutral, happy, sad and anger from male and female speech. Two elicited mode database were created and analyzed. The stated recognition accuracies for automatic emotion recognition from speech by using features like MFCCs, LPCs, F0, and Formants are in the range of 35-70% with a large combination of different features. In our approach we could successfully reduce the feature vector size and obtain overall recognition accuracies of 72.055% in male and 65.5% in female speech respectively.

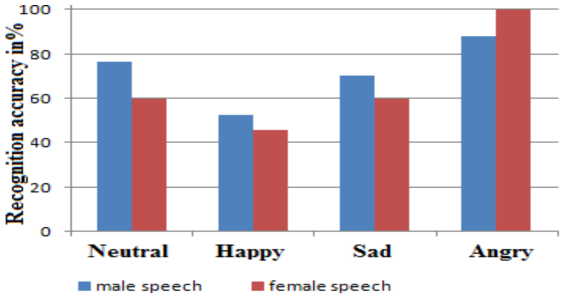

Fig 2 : Recognition accuracy obtained for male and female speech

REFERENCES

[1]. Banziger and K.R. Scherer, *"The Role of Intonation in Emotional Expressions," in Speech Communication*, vol. 46, pp. 252-267, 2005.I.Daubechis "Orthonormal Bases of Compactly supported wavelets" Communication on pure and Applied Math.Vol.41, 909-996. 1988

[2]. L.M.Bruce, C.H.Koger,J.Li, *"Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction"*, IEEE Transactions on geosciences and remote sensing, vol.40, No.10,October, 2002

[3]. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, *"Emotion recognition in human-computer interaction," IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[4]. S.A. Mallat. *"A Theory for Multiresolution Signal Decomposition: The wavelet Representation".* IEEE Transactions on Pattern Analysis And Machine Intelligence, 674-693, Vol.11 1989

[5]. I.Daubechis *"Orthonormal Bases of Compactly supported wavelets",* communication on pure and Applied Math.Vol.41, 909-996, 1988

[6]. S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 11;No.7, pp. 674-693, 1989

[7]. George Tzanetakis, Georg Essl, Perry Cook "*Audio Analysis using the Discrete Wavelet Transform"* *Computer Science Department *also Music DepartmentPrinceton35 Olden Street, Princeton NJ 08544.

[8]. LiMin Fu Neural Networks In Computer Intelligence Tata McGraw-Hill publishing company limited ISBN 0-07-053282-6 2003.

[9]. S. Rajasekeran, G. A. Vijayalakshmi Pai Neural Networks, Fuzzy Logic, and Genetic Algorithms, Synthesis and Applications, , Prentice-Hall of India, ISBN 8L-203-2L86-3 2004