International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010
1793-8201

Discrete Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition

Firoz Shah. A, Raji Sukumar. A and Babu Anto. P

***Abstract*—Automatic Emotion Recognition (AER) from speech finds greater significance in better man machine interfaces and robotics. Speech emotion based studies closely related to the databases used for the analysis. We have created and analyzed three emotional speech databases. Discrete Wavelet Transformation (DWT) was used for the feature extraction and Artificial Neural Network (ANN) was used for pattern classification. We can find that recognition accuracies vary with the type of database used. Daubechies type of mother wavelet was used for the experiment. Overall recognition accuracies of 72.05 %, 66.05%, and 71.25% could be obtained for male, female and combined male and female databases respectively.**

***Index Terms*— Automatic Emotion Recognition, Artificial Neural Networks, Affective Computing, Discrete Wavelet Transform**

I. Introduction

The interface between man and machine will become more meaningful if the machines can recognize the emotional contents. Emotions are the backbone of human interactions and are closely related to rational thinking perception, cognition and decision makings [1, 2]. Emotional cues can be analyzed from speech, facial expressions and gestures. In this work we are focusing on recognizing emotions from speech. Theories for emotional standards are mainly classified to two. One deals with discrete approach and the other deals with dimensional approach. The discrete approach related to universal basic emotions where as dimensional approach characterizes and distinguishes different emotions [3]. Since speech is the primary medium for interaction, speech based emotional studies are more significant. Emotions in speech do not alter the linguistic contents of speech but changes its effectiveness. Automatic emotion recognition systems finds applications in Human-Computer Interfaces (HCIs), humanoid robotics, text-to speech synthesis systems, forensics, lie detection, interactive voice response systems etc. Emotion recognition is a complex pattern recognition problem which relates with both cognitive and neural approaches [4, 5]. This paper is organized as follows; first we introduce the databases used. Next we have described our feature extraction procedure. In section IV we have introduced our pattern classification technique; in section V we stated the experiments and results, in section VI we added the discussions and results, with section VII we conclude the paper.

II. Emotional Speech Databases

Three elicited emotional speech databases were created for this experiment. Malayalam (one of the south Indian languages) was used for the experiment. Speakers under the age group of 30 years were used for creating the speech corpus. 10 male speakers and 10 female speakers participated in database creation. First database created consists of 300 male speech samples and the second database consists of 300 female speech samples. Third data set contains 600 samples of both male and female speech. The emotional speech database and their IPA format are given in Table I.

TABLE I : Formats of Words Used to Create The Emotional Speech Databases

Words in Malayalam	Words in English	IPA format
അമ്മേ	amme	//æ/m/ m/ æ//
അച്ഛാ	acha	//æ///tʃʰ/ ɑ ː //
മോളേ	mole	// m/ ɒ/ l/ ɛ//
മോനേ	mone	// m/ ɒ/ n/ ɛ//
എടാ	eda	// ɛ/ d/ɑ ː //
ലെതേ	lethe	// l/ ɛ// θ/ ɛ//
ദേവീ	devi	// d/ ɛ/ v/ ɪ//
ഞാനോ	njano	// n/ dʒ/ɑ ː / n/ ɒ//
കുട്ടീ	kutty	//k/ʊ/t/t/i//
മായേ	maye	// m/ ɑ ː / j/ ɛ//
അയ്യോ	ayyo	//æ/aɪ/ɒ//
ചേട്ടാ	chetta	//tʃʰ/t (ʰ)/ɑ ː //
വേണ്ട	venda	// v/ i ː / n/ d/ɑ ː //
കണ്ടു	kandu	// k (ʰ)/ɔ ː / n/ d/ ju ː //
പോയീ	poyi	// p/ ɒ/ i//
പോടാ	poda	//p/o/d/a://

പോടീ	pode	//p/o/d/I//
എടീ	ede	// ɛ/ d/i://
വാവേ	vave	//v/a:/v/ æ//
നീയോ	neeyo	//n/ ɛ/ ɔ ı/ ɒ//

III. Feature Extraction by Using Discrete Wavelet Transforms

Discrete Wavelet Transforms (DWTs) are orthogonal functions which can be implemented through digital filtering techniques and are basically originates from Gabor wavelets. Wavelets have energy concentrations in time and are useful for the analysis of transient signals such as speech signals. DWT is the most promising mathematical transformation which provides both the time –frequency information of the signal and is computed by successive low pass filtering and high pass filtering to construct a multi resolution time-frequency plane [6]. In DWT a discrete signal x[k] is filtered by using a high pass filter and a low pass filter, which will separate the signals to high frequency and low frequency components. To reduce the number of samples in the resultant output we apply a down sampling factor of ↓2.

The Discrete Wavelet Transform is defined by the following equation.

$$W(j,k) = \sum_j \sum_k X(k) 2^{-j/2} \psi(2^{-j} n - k) \quad (1)$$

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

The digital filtering technique can be expressed by the following equations

$$Y_{high}[k] = \sum nX[n]g[2k-1] \quad (2)$$

$$Y_{low}[k] = \sum nX[n]h[2k-1] \quad (3)$$

Where Y high and Y low are the outputs of the high pass and low pass filters

IV. Artificial Neural Networks

Artificial Neural Network (ANN) is an efficient pattern recognition mechanism which simulates the neural information processing of human brain. The ANN processes information in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and non linear units. The computational intelligence of neural networks is made up of their processing units, characteristics and ability to learn. During learning the system parameters of NN vary over time and are characterized by their ability of local and parallel computation, simplicity and regularity [7]. Multi Layer Perceptron(MLP) architecture is used for pattern classification in this work. The MLP architecture consists of one or more hidden layers. A signal is transmitted in the one direction from the input to the output and therefore this architecture is called feed forward. The MLP networks are learned with using the Backward Propagation algorithm and is widely using in machine learning applications [8].MLP uses hidden layers to classify successfully the patterns into different classes. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs [9].

V. Experimental Work and Results

Three experiments were done to evaluate the recognition accuracies of the four different emotions from speech viz neutral, happy, sad and anger. We have used a high quality studio recording microphone 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 capturing 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 mother wavelet. By using Daubechies-8 wavelet we performed the successive decomposition of the speech signals to obtain a good feature vector. The databases were divided in to two for training and testing of the classifier respectively. We used a proportion of 80% for training and remaining 20% for testing of the classifier in all the experiments.

In the first experiment, we have analyzed the male speech database consisting of 300 utterances. During testing of the classifier 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% with 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 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 obtained 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. The confusion matrix for male emotional speech database indicating the recognition accuracies of different emotions is given in Table II.

TABLE II : Confusion Matrix Obtained in The Case of Male Speech Database

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%

In the second experiment, we analyzed the female speech database consisting of 300 speech utterances. In analyzing the female speech the machine recognized 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 of 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 other emotions. An overall recognition accuracy of 66.5% could be achieved from this experiment. A confusion matrix indicating the recognition accuracies for different emotions from female speech are given in Table III.

TABLE III : CONFUSION MATRIX OBTAINED IN THE CASE OF FEMALE SPEECH DATABASE

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%

In the third experiment we have used the male and female speech combined database consisting of 600 utterances. During testing for recognizing the emotion neutral, machine could achieve a recognition accuracy of 75% and faced a confusion of 10% with both the emotions happy and sad, and the machine faced a confusion of 5% with the emotion anger. In the case of recognizing the emotion happy we could have obtain a recognition accuracy of 50%, and faced a confusion of 25% with the emotion neutral,20% confusion with the emotion sad and faced a confusion of 5% with the emotion anger. In recognizing the emotion sad we have obtained a recognition accuracy of 70% .and the machine faced a confusion of 15% with the emotion neutral, 10% with the emotion happy and a confusion of 5% with the emotion anger. In recognizing the emotion anger the machine obtained a recognition accuracy of 90%, and faced a confusion of 5% with both neutral and sad. There obtained no more confusion with the emotion sad. An overall recognition accuracy of 71.25% is obtained from this experiment. A confusion matrix indicating the recognition accuracies obtained for different emotions is given in table IV.

TABLE IV : CONFUSION MATRIX OBTAINED IN THE CASE OF MALE& FEMALE SPEECH DATABASE

Emotional Class	Neutral	Happy	Sad	Anger
Neutral	75%	10%	10%	5%
Happy	25%	50%	20%	5%
Sad	15%	10%	70%	5%
Anger	5%	0%	5%	90%

VI. RESULTS AND DISCUSSIONS

By using the three databases we have obtained the following recognition accuracies by using the same feature extraction methods and classification techniques. For the emotion neutral a recognition accuracy of 76.47% has been achieved in case of male speech database, 60% in case of female speech database and 75% recognition in case of combined male and female speech database. For recognizing the emotion happy machine could achieve a recognition accuracy of 52.94% in case of male speech database,46% in case of female speech database and 50% in case of male and female database. In recognize the emotion sad the male speech database can achieve a recognition accuracy of 70.58%, in case of female speech database we could achieve 60% and in case of combined male and female database recognition of 70% recognition could be achieved. In recognizing the emotion anger we can achieve a recognition accuracy of 88.23% in case of male speech, 100% recognition accuracy in case of female speech and 90% recognition accuracy in case of both male and female speech databases. The obtained recognition accuracies are given in Table V.

TABLE V : THE RECOGNITION ACCURACIES OBTAINED IN CASE OF THE THREE DATABASES FOR THE FOUR EMOTIONS.

Databases used	Neutral	Happy	Sad	Anger
Male	76.47%	52.94%	70.58%	88.23%
Female	60%	46%	60%	100%
Male& Female	75%	50%	70%	90%

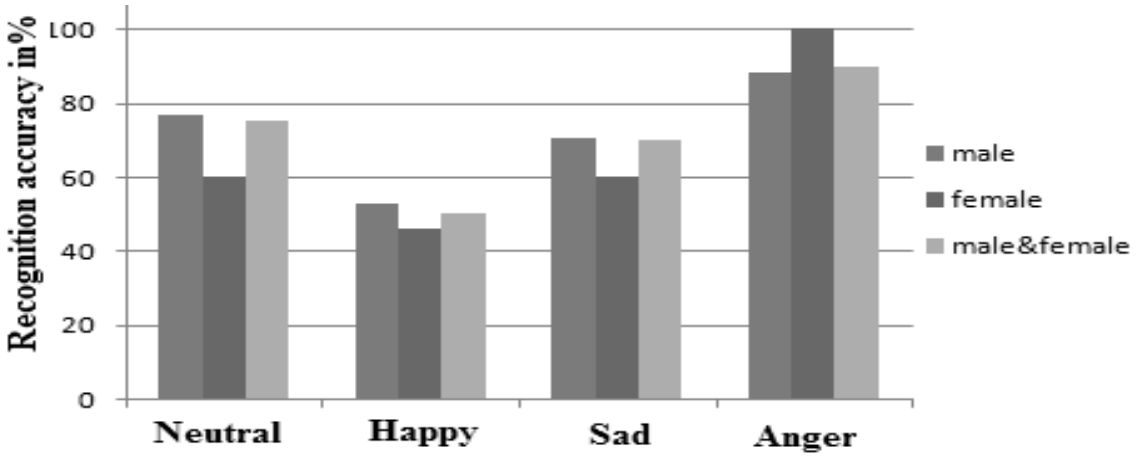

Figure 1: The recognition accuracies obtained for different emotions by using different databases.

VII. Conclusion

The recognition performance of different emotions from speech by using gender dependent and gender independent databases are carried out in this experiment. Artificial Neural Network was used for the machine learning purpose. Percentage recognition obtained for each emotions in the case of the different databases are compared. The results obtained from the experiments shows that emotion recognition from speech strictly depends on the databases used. The performance of the algorithm can be evaluated by using different databases.

References

[1] R.W.Picard Affective Computing. MIT Media Lab Perceptual Computing Section tech.rep., No.321 1995
[2] Petrushin,V., Emotions in speech: Recognition and Application to Call Centers Artificial Neural Network Engineering Nov 1999
[3] B.Vlasenko, B.Schuller, A.Wendemuth, and G.Rigoll Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In Proceedings of Affective computing and Intelligent Interaction 2007
[4] Tao J.H.kang Y.G. Features importance analysis for emotional speech classification, In proceedings of lecture notes in computer science 3784Springer 2005
[5] Pantic M. ,Rothkrantz, L.J.M., Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol.91, No.9., 1370-1390 2003
[6] S.A. Mallat: A Theory for Multi resolution Signal Decomposition The wavelet Representation. IEEE Transactions on Pattern Analysis And Machine Intelligence, 674-693, Vol.11 1989
[7] Haykin, S., , Neural networks: A comprehensive foundation, Englewood Cliffs, NJ: Prentice-Hall, New York1999
[8] Limin Fu Neural Networks in Computer Intelligence: Tata McGraw-Hill New Delhi India 2003
[9] Bishop Christopher. Neural networks for Pattern Recognition, Oxford University Press 1995

Firoz Shah A is a Research scholar working in the area of emotional speech processing at School of Information Science and Technology Kannur University, Kannur District, Kerala State, India. He has received MSc. Degree in Electronics from Mahatma Gandhi University, Kerala. His main research interest includes Emotional speech processing, pattern classification, Artificial Intelligence and Signal Processing.

Raji Sukumar.A is a Research scholar working in the area of speech and language processing at School of Information Science and Technology Kannur University, Kannur District, Kerala State, India. She has received MSc. Degree in Computer Science from Kannur University, Kerala and MCA Degree from Indira Gandhi National Open University New Delhi. Her research interest includes Speech processing, Artificial Intelligence, Natural language processing.

Babu Anto P is a Reader in information Science and Technology at the Department of Information Science and Technology, School of Information Science and Technology, Kannur University, Kerala State, India. He holds his MSc. and Ph.D. Degrees in Electronics from CUSAT (Cochin University of Science and Technology), Kerala, India. His current research interests include Speech and Emotion recognition, Data Mining, Speaker Recognition, and Visual Cryptography. He has published research papers in reputed national and international journals.