International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **06** Special Issue **01** | **2021**

Framework for detection of Emotion using Speech Signal

[1]Moiz A.Hussain
Electrical Engineering Dept, Dr.V.B.Kolte C.O.E, Malkapur, Dist. Buldana. (M.S.)

ABSTRACT

This paper reports the results of detecting emotions from speech signals, with particular focus on extraction emotion from short speech utterances. The main focus is on distinguishing the differences between anger and neutral speech. To obtain the feature vectors, different methods are used, such as: Support Vector Machines (SVM), Multilayer Perceptron (MLP), Generalized feed forward (GFF). The database used is from the 'Berlin Database of Emotional Speech'. The database contains 10 different sentences, spoken by 10 speakers (5 female, 5 male) in 7 different emotions (Neutral, Anger, Boredom, Disgust, Anxiety/Fear, Happiness, Sadness) in German. Audio files are in WAV format: 16 kHz, 16 bit, mono.

Keyword Speech Signal, SVM, MLP, GFF

1. INTRODUCTION

As we begin the 21st century where the implementation of computers in modern industry is widespread, the focus on computer and human interaction is increasing. Human interaction is said to be: 'speech, eye contact & gesture etc. with speech being the most common form of communication. With more effective computer and human interaction, in particular the recognition of different emotions from speech, greater efficiency would be achieved in communications.

Anger being recognised as the most important emotion to detect in computer to human interaction, as the result may lead to an end in interaction between the human and computer. Petrushin [10], stated anger as the most important emotion in business. Anger is a term for the emotional aspect of aggression, as a basic aspect of the stress response in animals whereby a perceived aggravating stimulus 'provokes' a counter response which is likewise aggravating and threatening of violence.

Emotion Recognition

The solutions to emotion recognition depend on:

 Which emotion should be recognised

 What purpose the emotion should be recognised for Emotion recognition has applications in 'talking' toys, video and computer games, and call centres. Automatic emotion recognition of speech can be viewed as a pattern recognition problem. The results produced by different experiments are characterised by:

 the features that are believed to be correlated with the speaker's emotional state,

 the type of emotions that humans are interested in;

 the database used for training and testing the classifier;

 the type of classifier used in the experiments.

To compare classification results, the same dataset must be used and there must be an agreement on the set of emotions.

Database

The performance of an emotion classifier relies heavily on the quality of the database used for training and testing and its similarity to real world samples (generalization). In order to conduct the experiment of recognizing human emotion, an audio database that can convey the emotional state of human was used. To ensure the diversity of the database, we used audio samples from ten subjects, speaking ten sentences spoken by five (5) male & five (5) female for seven classes: *neutral, anger, boredom, disgust, fear, happiness, and sadness*. The Berlin Database of Emotional Speech was used for this purpose [4].In this approach, the speech signal was re-sampled to 10 kHz, and the silence segments at the beginning and the end of the speech were cut out artificially. Then the whole database was divided into two parts for the purpose training & testing.

Computer Simulation Experiment Feature Extraction

The author have developed a program in MATLAB to obtain statistical parameters namely formant frequencies, entropy, variance, minima, median, LPC of a sound. Thus dataset for all 500 speech samples is prepared to feed to Neural Network for emotion recognition

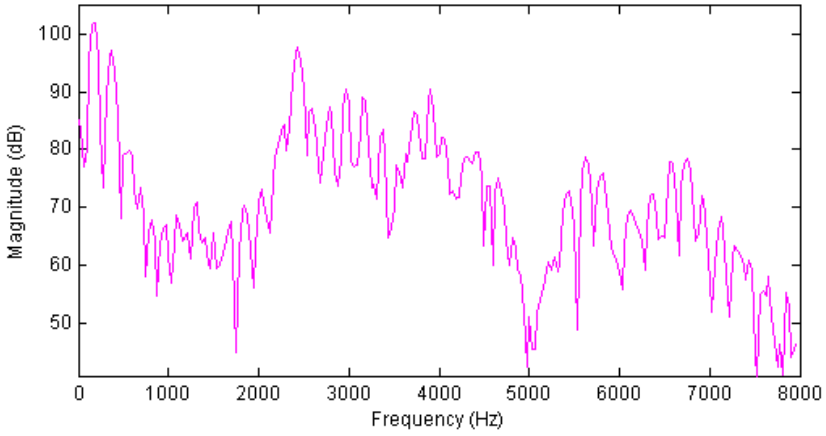

Figure 1 LPC Spectra of speech signal

Pre-Processing

Filtering to remove noise and normalisation to place all the sounds from the database into the same dynamic range. Segmentation into 20ms blocks.

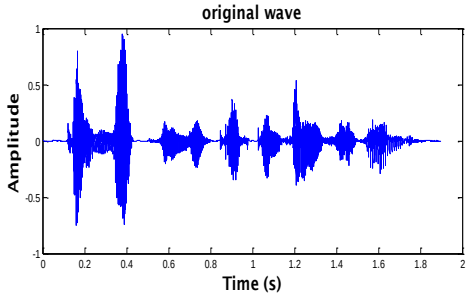

Figure 2 - Speech Signal

Speech Activity Detection

The speech activity detector is a self-normalising, energy based detector that tracks the noise floor of the signal and can adapt to changing noise conditions, used to remove silence from the segments.

Speech Analysis

Energy Features: Using 5 types of energy contours, the, ‘Log Entropy’, 'Shannon Energy', ‘Threshold Entropy’, ‘Sure Entropy’, ‘Norm Entropy’

Formant Features: Using 5 types of formant frequency contours, ‘Formant0’, ‘Formant1’, ‘Formant2’, ‘Formant3’, and ‘Formant4’

Audible Duration Features: Audible segments are determined by choosing a threshold below the maximum energy, and then a contour is produced showing audible/inaudible segments of speech.

Calculation of Feature Vectors

The main areas of focus for the feature vectors are:

- pitch related (fundamental frequency)
- loudness (energy)
- segments (audible duration)

Energy and Pitch Contours

From the two feature types (energy and pitch), calculation of maximum, minimum, mean, standard deviation, shimmer/jitter, first derivative.

Clustering of Feature Vectors

These feature vectors will then be used to make a clustered algorithm, to determine if a speech signal contains feature vectors relating to anger speech, or neutral speech.

Generalized Feed Forward Neural Network (GFFNN)

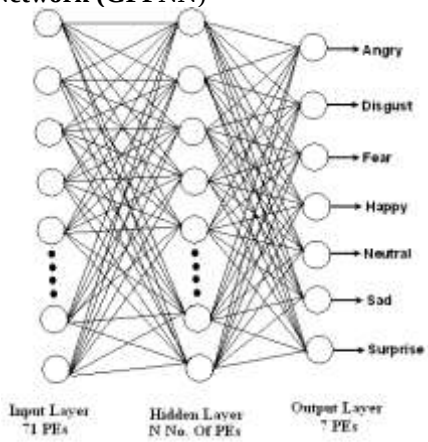

Figure 3: Schematic Diagram of GFFNN

Generalized feed forward networks are a generalization of the MLP such that connections can jump over one or more layers. In theory, a MLP can solve any problem that a generalized feed forward network can solve. In practice, however, generalized feed forward networks often solve the problem much more efficiently. A classic example of this is the two spiral problem. Without describing the problem, it suffices to say that a standard MLP requires hundreds of times more training epochs than the generalized feed forward network containing the same number of processing elements.

Neural Network for Emotion Recognition:

The generalized procedure for emotion recognition from speech signal using different feature extraction techniques is shown in figure4. We have used LPC, Variance, Formant, Entropy, Median, Minima for feature extractions and SVM, MLP, GFF for emotion recognition.

Table 1: SVM recognition results of training dataset

Output / Desired	On	Oa
On	5	0
Oa	0	15

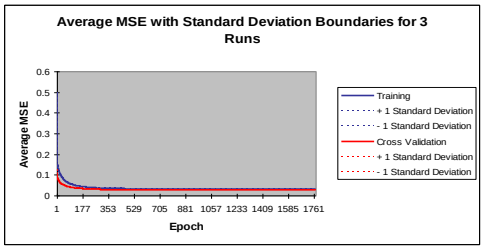

Figure 4: Structure of Emotion Recognition System

The structural components of the sex -independent emotion recognition system are depicted in Figure 3. It consists of four modules: speech input, feature extraction, neural network for classification, and the recognized emotion output.

Results

The SVM, MLP & GFF classifier was used to test the proposed feature vector of speech. The Train N Times training method was used in all the classifier to train the neural network, and experimental results were obtained by using 10% cross-validation (C.V.) data. The recognition result for SVM, MLP & GFF classifier for both training & C.V. dataset are shown in table below.

Table 2: SVM recognition results of cross validation dataset.

Output / Desired	On	Oa
On	71	0
Oa	0	112

Table 3: SVM recognition results of cross validation dataset based on the performance.

Performance	On	Oa
MSE	0.044861309	0.044839532
NMSE	0.188928621	0.188836907
MAE	0.163861831	0.164508908
Min Abs Error	8.95074E-05	8.4735E-05
Max Abs Error	0.575750136	0.574266207
r	0.962992493	0.963260712
Percent Correct	100	100

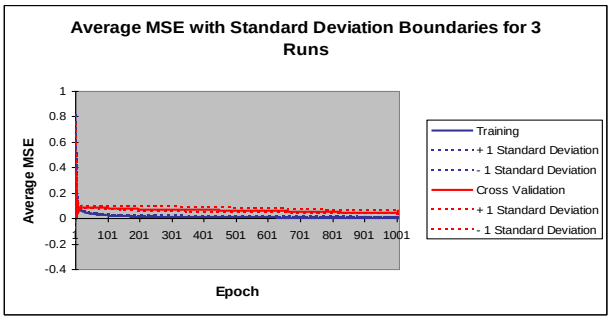

Figure 5: Average Mean Square error of Cross validation dataset & Training dataset.

Table 4: MLP recognition results of cross validation dataset.

Output / Desired	On	Oa
On	0	0
Oa	5	15

Table 4: GFF recognition results of cross validation dataset based on the performance

Performance	On	Oa
MSE	0.190064333	0.326545268
NMSE	0.782718204	1.344770595
MAE	0.245120223	0.453358326
Min Abs Error	0.002068037	0.001266136
Max Abs Error	1.055142814	1.028649687
r	0.65412139	0.278162011
Percent Correct	68.42105263	96.26168224

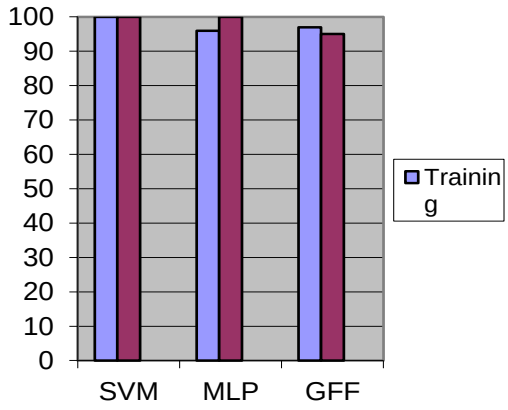

Figure 7: Comparison of recognized emotions by SVM, MLP &GFF Neural Network

Conclusion

In the work conducted for this project, different methods to distinguish the difference of anger speech and neutral speech were explored. From these results, several conclusions can be drawn. First, decoding of emotions in speech is complex process that is influenced by cultural, social, and intellectual characteristics of subjects. People are not perfect in decoding even such dominant emotions as anger and happiness. Second, anger has different variations, (hot anger, cold anger, etc.) that have different acoustic features and will dramatically effect the accuracy of recognition. It is recommended that these variations be taken into account when labelling a speech database.

References

[1] Klaus R. Scherer, Vocal communication of emotion: A review of research paradigms, Speech Communication 40, pp. 227-256

[2] Lili Cai, Chunhui Jiang, Zhiping Wang, Li Zhao, Cairong Zou, "A Method Combining The Global And Time Series Structure Features For Emotion Recognition In Speech", IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, December 14-17, 2003 0-7803-7702-8/03/ 2003 IEEE.

[3] Yi-Lin, Gang Wei, "Speech Emotion Recognition Based on HMM and SVM", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005 0-7803-9091-1/05/2005 IEEE.

[4] Muhammad Waqas Bhatti[1], Yongjin Wang[2] and Ling Guan3, "A Neural Network approach for Human Emotion Recognition in Speech", 0-7803-8251-X/04/2004 IEEE.

[5] Felix Burkhardt, Miriam Kienast, Astrid Paeschke and Benjamin Weiss, "Berlin Database of Emotional Speech" available at http://pascal.kgw.tu-berlin.de/emodb/

[6]Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., "Emotion recognition in human-computer interaction", IEEE Signal Processing magazine, Vol. 18, No. 1, pp. 32-80, Jan. 2001.

[7] J. Nicholson, K. Takabashi and R. Nakatsu, "Emotion Recognition in Speech Using Neural Network", Neural Information Processing, 1999.

[8] Li Zhuo, Xiungmin Qiun, Cuirong Zou, Zhenyung Wu, "A Study on Emotional Feature Analysis and Recognition in Speech Signal" Journal of China Institute of Communications, Vo1.21, No.10, pp18-25, 2000.

[9] François Thibault," Formant Trajectory Detection Using Hidden Markov Models", Special Project Course Report MUMT 609, December 14 2003.

[10] Petrushin, V., "Emotion in Speech: Recognition and Application to Call Centers", *in Proc. of Artificial Neural Networks in Engineering*, pp. 7-10, Nov. 1999.