Enhanced HMM Speech Emotion Recognition using SVM and Neural Classifier

Preeti Suri
M.tech (CSE)
IGEF – Abhipur
Punjab Technical University
Jalandhar, India

Bhupinder Singh
H.O.D. (CSE)
IGEF - Abhipur
Punjab Technical University
Jalandhar, India

ABSTRACT
Emotion classification has been a research area for decades. Identifying the correct emotion of the user helps in a lot of areas like crime investigation and other related research works. This particular thesis work has been done using 4 emotion categories in which SVM and Custom Neural Network has been used as a classifier. This thesis work is consisted of two main sections. The first section is called the training part and the second part is called the testing part. In the training part we have used different voice files of different categories like happy, sad, angry and aggressive to train the system according to the classified properties of the speech samples. To train the system feed forward method has been used and the database format is .mat. In the testing section we have used SVM to binarize the features of the input sample and Neural to finally classify the entire architecture. The custom neural network in this research work has been provided two categories of sample, the first sample is the training data set and the final sample is the testing data set. The finally accuracy of the plot comes out to be more than 90 %

General Terms
Speech Emotion Recognition, HMM, SVM, Neural Networks

Keywords
Emotion States, Energy, Pitch, Emotion Classifier

1. INTRODUCTION
Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or his mental state to others. It is a channel of human psychological description of one's feelings. Emotions are a key part of speech [1]. Automatically detecting emotion in a recording can enhance human computer interaction. It also enables various other kinds of analyses, such as search for paralinguistic phenomena, the honesty of the speaker, etc.

Automatic Speech Emotion Recognition is a very active research topic in the Human Computer Interaction (HCI) field and has a wide range of applications. It can be used for in-car board system where information of the mental state of the driver maybe provided to initiate his/her safety. In automatic remote call center, it is used to timely detect customers‟ dissatisfaction. In E-learning field, identifying students‟ emotion timely and making appropriate treatment can enhance the quality of teaching [1]. Now days, the teachers and students are usually separated in the space and time in E-learning circumstance, which may lead to the lack of emotional exchanges. And the teacher cannot adjust his/her teaching method and content according to the students‟ emotion. For example, when there is an online group discussion, if students are interested in the topic, they will be lively and active, and show their positive emotion. On the contrary, if they get in trouble or are not interested in it, they will show the opposite emotion. If we detect the emotion data, and give helpful feedback to the teacher, it will help the teacher to adjust the teaching plan and improve the learning efficiency [2].

In this paper various features are extracted from each utterance for the computational mapping between emotions and speech patterns. The selected features are then used for training and testing a modular neural network. Classification result of neural network and SVM classifiers are investigated for the purpose of comparative studies.

2. SPEECH DATABASE
In our proposed work we have used speech samples for the database. In the database we find properties of the speech signals and then we store them into the database. The question comes that how we are going to store hundreds of files in the database. The procedure would be as follows. First of all we would fetch the properties of the voice samples. All those properties which are required would be computed and then it would be stored into an array. The array would move on as the files would move. We would fetch the features and would take the average by the end and then store them into the database for each category of the voice which we have taken i.e. **HAPPY, SAD, ANGRY AND AGGRESSIVE.**

2.1 Voice Files
The voice files are the files which would be processed for the feature extraction.

2.2 Properties
When we would process the voice files their properties would be fetched. For the feature extraction there are several algorithms which can be used. In our approach we have used HMM algorithm for the training purpose.

3. SPEECH EMOTION RECOGNITION SYSTEM
Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on speech signal, extracting some features which contain emotional information from the speaker’s voice, and taking appropriate pattern recognition methods to identify emotional states [3]. Speech emotion recognition system contains four main modules: speech input, feature extraction,

SVM based classification with neural network, and emotion output.

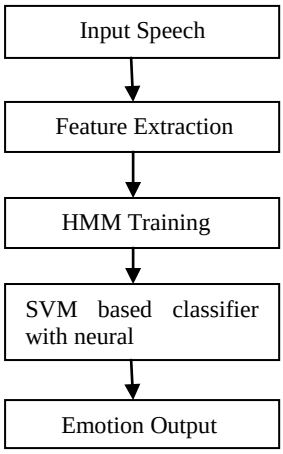

Fig 1: Speech Emotion Recognition System

4. RELATED WORK

There are two sections in our research work. The sections are explained as follows.

4.1 Training

In the training section we would be taking fifty voice samples of each and every category taken for the classification. In this scenario, we would be fetching properties of each voice sample and after putting them into an array. We would be storing the average of each property of each section into the database. To achieve this particular task, we would be using HMM algorithm. The training section ensures that the database gets trained properly so that at the time of testing it produces extensive results. The features of training are as follows

4.1.1 Maximum Frequency

The maximum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the maximum peak is called the maximum frequency of the voice sample.

4.1.2 Minimum Frequency

The minimum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the minimum peak is called the minimum frequency of the voice sample.

4.1.3 Average Frequency

The average frequency can be calculated using two techniques. The first technique is to add all the frequency samples and then divide the entire sum with the total number of frequency. The second method is a very ethical method in which we can add the minimum frequency and the maximum frequency and then we can divide them by two.

Avg Frequency = (Minimum frequency + Maximum Frequency)/2

4.1.4 Spectral Roll off

The spectral roll off in terms of development can be said as the difference between the maximum frequency differences with the adjacent frequency. The position of the frequency (max) can be stored into an array and similar of the adjacent node and then the difference can be calculated.

4.1.5 Noise Level

Ethically the noise level is the extra number of bits which has been added into the voice sample. If the noise is uniform then the noise level can be calculated by taking the difference of each frequency sample and the threshold of the voice sample. There are two categories of noise level: uniform noise and non uniform noise.

UNIFORM NOISE: Uniform noise is the noise which is simultaneously same all over the voice sample.

NON UNIFORM NOISE: The non uniform noise does not remain constant all over the sample.

4.1.6 Pitch

It is the average value of the entire voice sample.

4.1.7 Spectral Frequency

The spectral frequency is the frequency of the voice pitch next to the highest voice sample.

4.2 Testing

At the time of testing we would be using a combinational algorithm using the SVM and NEURAL feed forward method. In this part, we would be binarizing the saved data of the database and would provide it to the NEURAL classifier. On the basis of the saved database, the neural classifier would match the properties with the uploaded file and would produce a result. The classification would be done on the basis of 4 categories: happy, sad, angry and aggressive.

5. FEATURE EXTRACTION ALGORITHM

5.1 HMM

HMM stands for HARCOV'S META MODEL. It is a worldwide known algorithm for the training of the data set. It extracts the features of the voice sample and saves them to the database for the future use. The maximum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the maximum peak is called the maximum frequency of the voice sample. It is viewed as the counter part of the training and it is used to sample size the data for the further processing. In this approach we take each sample of data set as a unique item which has to be processed [5].

The HMM consist of the first order markov chain whose states are hidden from the observer therefore the internal behavior of the model remains hidden. The hidden states of the model capture the temporal structure of the data.

HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. During classification, a speech signal is taken and the probability for each speech signal provided to the model is calculated.

5.2 SVM

SVM stands for stands for support vector machine. It is unique algorithm in itself. The SVM changes the entire input into some sort of binary form so that the matching accuracy may increase. As we all know that even if we take the input voice of only one person, it can't same each and every time in terms of the frequency. Now if we binarized each and every thing the matching percentage may increase as if eighty percent of the data even matched, it would be sufficient enough to produce a result.

5.3 Neural

Neural Networks is one of the most advanced classifiers in the testing category. The neural has a feed forward method. The feed forward method takes one input as a training sample and another input as the target sample. The input sample is the data stored in the database on the basis of all the features which have been extracted at the time of training.

If P is the input sample then P would be defined as
P =sum (all. Features (input));

In the same manner, there would be the testing feature or the target sample. The target sample would be the scenario which would be fixed and would be provided as an input.

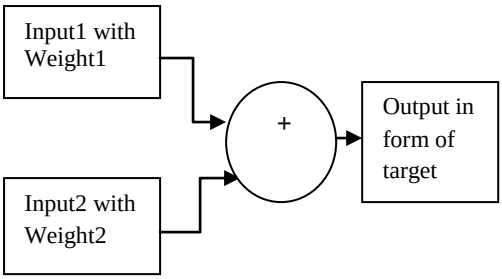

Fig 2: General working principle of the neural networks

Here the central block is termed as the **NEURAL CLASSIFIER**. There are two input samples where the neural classifier generates the weight accordingly for the first input which has been taken from the database. The second input is the target set which is to be tested. The neural takes each sample as a neuron and explains to the architecture that how the input is going to react and how the result is going to be proceeded. Finally it produces a binary result. If we proceeds more than one sample category, the neural will have be combined.

6. EXPERIMENTAL RESULTS

Table 1. Sad Matrix

Emotion	Sad	Others
Sad	80	13.3333
Others	20	86.6667

The above tabular form represents the accuracy for the sad category in which 80 is the true positive and 20 is the false positive for the same. Where as if we look at with the other variances in the same contrast, we would find that 86.66 % is the false positive of the other categories we take sad file as an input.

Table 2. Angry Matrix

Emotion	Angry	Others
Angry	73	12.1667
Others	27	87.8333

The above tabular form represents the accuracy for the angry category in which 73 is the true positive and 27 is the false positive for the same. Where as if we look at with the other variances in the same contrast, we would find that 87.83 % is the false positive of the other categories we take angry file as an input.

Table 3. Overall Confusion Matrix

Emotion	Sad	Happy	Angry
Sad	80	10	4
Happy	6	70	15
Angry	15	7	73

The above confusion matrix represents the all the results. We can conclude that using Neural as a classifier we gets a classified category % age architecture for sad, happy and angry voice categories in which the significance level of sad is 80% out of 100 percent where as 70% is for happy and 73 % is for angry.

The calculation formula is as follows:
Accuracy percentage = (Tested File Category/ Total Files Uploaded)*100

7. CONCLUSION

Undergoing the estimated procedure, we expect our conclusion to be a better accurate system for the analysis of the audio files to detect the emotions in the field of clustering. We expect the accuracy to be increased by 2 to 5 percent in comparison with the ANN. SVM combined with HMM and neural is expected to work in better manner because the training set created with the help of SVM and HMM puts a strong emphasis in searching into the inner clusters of the files.

Although the results are efficient enough to provide solution, but our work provides a lot for the future researchers. The current scenario is not suitable for noisy audio files, if we increase the noise level of audio files the current scheme might fail to produce efficient results.

Future works of this work may involve the use of back propagation of neural network. In future, work can also be done to create more groups into the inner cluster of the files stored so that the searching becomes easy. The future parameters can add the time slots of the frequencies at which the frequencies are consistent.

8. ACKNOWLEDGMENTS

I would like to place on record my deep sense of gratitude to Mr. Bhupinder Singh, H.O.D of Computer Science Department of IGCE, Abhipur for his generous guidance, help and useful suggestions. And all faculty members of CSE department for attending my seminars and for their insightful

comments and constructive suggestions to improve the quality of this research work.

I am extremely thankful to Dr. Promila Kaushal Principal, IGCE Abhipur, for providing me infrastructural facilities to work in, without which this work would not have been possible.

9. REFERENCES

[1] http://www.sersc.org/journals/IJSH/vol6_no2_2012/15.pdf

[2]http://www.ffri.hr/~ibrdar/komunikacija/seminari/Ververidis,%202006%20%20Emotional%20speech%20recognition.pdf

[3] http://www.mmk.ei.tum.de/publ/pdf/02/02sch2.pdf

[4]http://www.ee.columbia.edu/~dpwe/e6820/proposals/kisang.pdf

[5] Xia Mao, Lijiang Chen, Liqin Fu, "Multi-level Speech Emotion Recognition Based on HMM and ANN", *2009 WRI World Congress, Computer Science and Information Engineering*, pp.225-229, March 2009.

[6]http://poseidon.csd.auth.gr/LAB_PEOPLE/Ververidis/Ververidis_ICASSP_2004.pdf

[7] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", *International Journal of Computer Applications*, vol.1, pp.6-9,February 2010.

[8] Liudong Xin "An Efficient Approach for Audio mining Reliability and Sensitivity Analysis" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: systems and humans, vol. 38, no. 1, January 2008.

[9] Chattopadhyay, S "Variable ordering for sharedHMM targeting node count and path length optimisation using particle swarm technique" Computers & Digital Technique, Volume: 6 Issue: 6.

[10]Frank Pfenning'' Lecture Notes on HMM 15-122: Principles of Imperative Computation'' Lecture 19 October 28, 2010.

[11] Bergman, Andre A. Cire, Willem-Jan van Hoeve, J.N.HookerTepper School of Business, Carnegie Mellon University 5000 Forbes Ave., Pittsburgh, PA 15213, U.S.A.

[12] Dayou Liu, Shengsheng Wang "k mean with Implied Literals: A New knowledge Compilation ApproachDavid".

[13] Manish Verma, Mauly Srivastava,"A Comparative Study of Various Clustering Algorithms in Data Mining", *International Journal of Engineering Research and Applications* Vol.2, Issues 3, May-Jun 2012, pp.1379-1384.

[14] Xavier Anguera, Member, IEEE, Simon Bozonnet, Student Member, IEEE, Nicholas Evans, Member, IEEE*,"* Speaker Diarization: A Review of Recent Research", *FIRST DRAFT SUBMITTED TO IEEE TASLP*: 31 MARCH 2010

[15] Indian Institute of Technology, Kanpur2LTI, School of Computer Science, Carnegie Mellon University, Pittsburgh3International Institute of Information Technology, Hyderabad

[16] International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974.

[17] S.B. ; Electronics Laboratory, General Electric "k mean " Computers, IEEE Transactions on (Volume:C-27 , Issue: 6),pp 509 - 516 .

[18]http://perso.telecomparistech.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page2679.pdf