An Automatic Emotion Recognizer using MFCCs and Hidden Markov Models

Chandni, Garima Vyas, Malay Kishore Dutta
Department of Electronics & Communication Engineering
Amity University, Noida, India
chandni.mittal025@gmail.com, gvyas@amity.edu
malaykishoredutta@gmail.com

Kamil Riha, Jiri Prinosil
Faculty of Electrical Engineering and Communication
BRNO University, Czech Republic
rihak@feec.vutbr.cz, prinosil@feec.vutbr.cz

***Abstract*: In this paper, the proficiency of continuous Hidden Markov Models to recognize emotions from speech signals has been investigated. Unlike the existing work which considers prosodic features for automatic emotion recognition, this work proposes the effectiveness of the phonetic features of speech particularly, Mel-Frequency Cepstral Coefficients which improves the accuracy with reduced feature set. The continuous speech emotional utterances used in this work have been taken from the SAVEE emotional corpus. The Hidden Markov Model Toolkit (HTK) version 3.4.1 was utilized for extraction of the acoustic features as well as generation of the models. Optimizing the acoustic and pre-processing parameters along with the number of states and transition probabilities of the Markov Models, the trials give us an average accuracy of 78% and highest accuracy of 91.25% for four emotions sadness, surprise, fear and disgust.**

Keywords: Emotion; recognition; HTK toolkit; Mel frequency cepstral coefficients.

I. INTRODUCTION

In recent years, great advances have been made in the area of recognition of emotions from speech. Apart from the facial expressions of the human beings, speech signal has been corroborated to be one of the most propitious modalities for automatic emotion recognition [12] as the speech signal does not only comprise of the word for word sense but also carry a wealth of supplementary data. The emotion recognizer utilizes this additional information from the speech signals to classify the emotions into different categories. Researches in the past have used a lot of different techniques to classify emotions from speech but mostly they utilized the common signal features such as pitch, intensity, spectral density, formants etc. such as in [1-5]. Yet, efficiency of such systems which make use of these prosodic features to differentiate emotions into four or more classes is low.

The key contribution of this paper is a method to accurately attach an emotional label to a continuous speech sample using the Hidden Markov Model Toolkit and Mel Frequency Cepstral Coefficients. In doing so the silence part in the speech was not considered which further enhanced the accuracy of the system. High efficiency was achieved by optimizing the acoustic parameters, number of states of the HMM for each emotion and the transitional probabilities between the states. The algorithm followed in this paper also demonstrates the usefulness of the Mel Frequency Cepstral Coefficients rather than the conventional prosodic features to recognize emotions from the speech. The HTK toolkit was used for emotional labelling [6-7] which is conventionally used for speech recognition. The future scope of the proposed technique is high because it is less complex than the existing algorithm but gives better efficiency.

The rest of the paper is organized as follows. Section II describes the methodology. Section III describes the database used in this paper. Section IV describes the results and discussion part, the conclusion and future work which may be possible in this area is in Section V.

II. METHODOLOGY

Automatic recognition of emotions from a speech is a very challenging task, as benefits of acoustic features in emotional classification is not clear. The work done in this paper suggests the use of Mel Frequency Cepstrum Coefficients (MFCCs) for acoustical analysis and then using Hidden Markov Models as classifiers. The Speech emotion recognition system consists of modalities such as the input speech signal, acoustical feature extraction, training of the HMMs and testing the capability of the classifier to assign emotional labels to the speech sample. The methodology used in this contribution consists of the following steps:

Step 1: Labelling of Emotional Database: Four emotions i.e. Surprise, Sadness, Fear and Disgust from the SAVEE database is split in the ratio of 2:1 for training and testing. The utterances were labelled with the help of *Wavesurfer* [8].

Step 2: Acoustical Analysis: The training and testing audio clips are transformed into a sequence of coefficient vectors extracted from the audio clips in the database.

Step 3: Defining the HMMs: Prototypes of Hidden Markov Models are created and initialized for each emotion.

Step 4: Training of the HMMs: The initialized models are trained with the training dataset.

Step 5: Task Definition: The task grammar is defined for the recognizer. Task Dictionary is also defined.

Step 6: Testing: Testing the recognizer with the test dataset has been done.

Step 7: Performance Evaluation: Performance of the automatic emotion recognizer built so far is evaluated. The algorithm followed in this paper demonstrates the usefulness of the Hidden Markov Models and the Mel frequency Cepstral Coefficients to recognize emotions from the Speech. The main block diagram is shown in Fig.1.

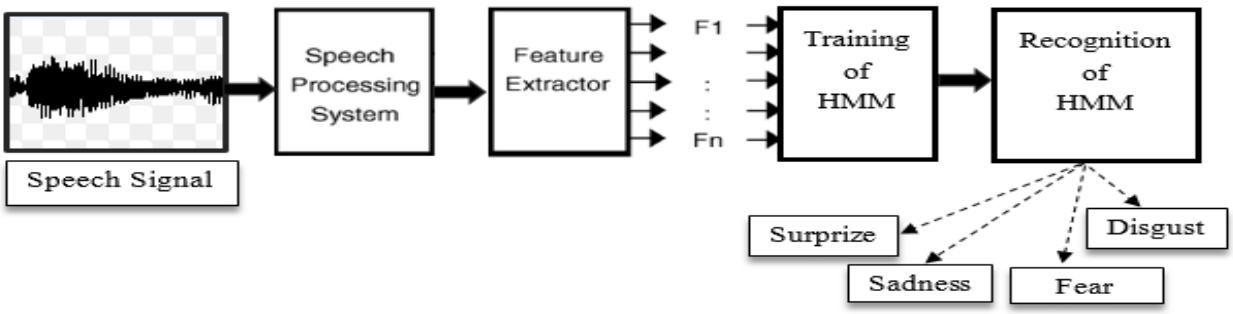

Fig. 1: Block Diagram of Emotion recognizer system

A. MFCCs

The Mel-frequency Cepstral Coefficients estimates the frequency response of the human auditory system hence they present a more reputable illustration of the speech signal. The speech samples of average 40 seconds length are segmented into frames with the help of Hamming Window function. It is defined with the following equation:

$$w(n) = 0.54 - 0.46\cos\frac{(2\pi n)}{(N-1)}, 0 \le n \le N-1 \ldots . (1)$$

The Mel Conversion takes place as written in the equation given below:

$$Mel(f) = 2595\log\frac{(1+f)}{700} \qquad \ldots\ldots\ldots\ldots\ldots (2)$$

A sequence of triangular Band-Pass filters used to pass the frequencies on the basis of Mel scale. After this, the obtained frequencies are again transformed back into a time domain by applying the discrete cosine transformation (DCT) on it [9], [11], [13]. The acoustical analysis in the Hidden Markov Model Toolkit is performed with the HCopy tool. The Fig. 2 depicts the complete process of conversion of audio clips into coefficient vectors. Here 13 MFCCs, 13 first differential of MFCCs (dMFCCs) and 13 double differential of MFCCs (ddMFCCs) vectors are utilized which makes a 39 dimensional network.

In this paper, the MFCCs were the chosen acoustical feature because of the reason that these are real numbers, hence they offer an excellent depiction of the spectral properties of the speech signal. Also, using only MFCC to extract emotions from the speech signal has the additional benefit that it is a self-sufficient methodology which does not have the need of calculating any other acoustical features. MFCCs are a standalone feature MFCCs are very compact and have ascertained to be much more proficient than other prosody based features [15], [16].

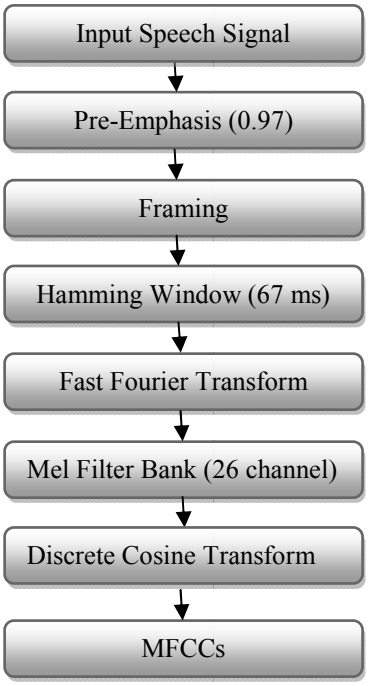

Fig. 2: MFCC Extraction Process in HTK

B. Training of HMMs

The Hidden Markov Models have been used for a very long time for speech recognition. [14], [15], [16]. But their applications in emotion recognition tasks are relatively less. Since speech is not stationary, the HMMs provide the flexibility to model voice as a sequence of states. So, the Hidden Markov Models were the choice of classifier in this contribution. Fig. 3 shows the HTK protocol to build HMM [10] based recognizers. The HTK tool HInit/ HCompV was used for initialization process of the prototype Hidden Markov Model and HRest/HERest was used to train the initialized model with the training dataset.

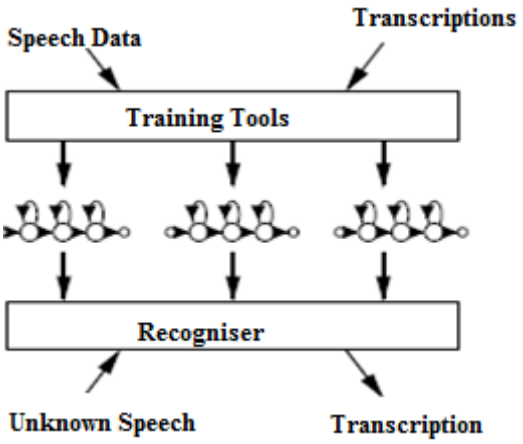

Fig.3: Building HMM through HTK

Creating an emotion recognition system in the Hidden Markov Model Toolkit happens in phases. At first, the training data and its corresponding transcripts are transformed into a HTK compatible format. Then these modified datasets are used to train the various HMM classifiers modelled for each emotion. At last the performances of the trained classifiers are tested against an unseen set of test data. The results thus obtained from the above process are analyzed with the help of tool HResults. The phases of HTK are enlightened in more detail in Fig 4. It describes the process of building a recognizer in HTK [6]. The same phases were followed in building a speaker independent, automatic speech emotion recognizer. The states of the HMM defined in this contribution are different for every emotion and hence optimum results were achieved from the recognizer. The desired acoustical features are MFCCs utilizing '*C*0' as the energy constituent. In this paper a thirty-nine dimension MFCC, delta and acceleration feature vector has been calculated for the 44.1 kHz sampled signals at 65 ms intervals.

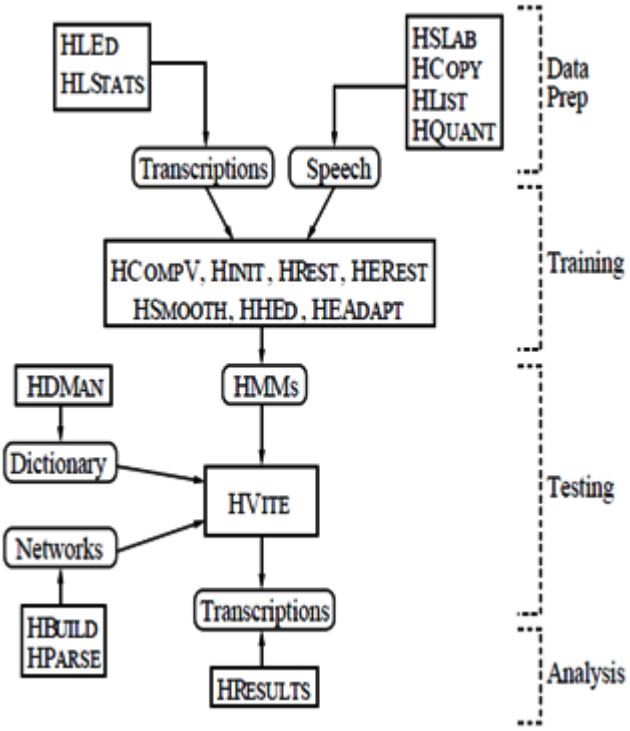

Fig.4: Phases of Building a recognizer in HTK

The HTK tool 'HCopy' was used for enumeration of the MFCCs. The output of this tool was kept in a compacted format; in addition, a cyclic redundancy checksum was also added. The 39 dimensional feature vector thus obtained comprised of 13 MFCCs, their first derivatives as well as their second derivatives. A part of the MFCC, dMFCC and ddMFCC vectors obtained for the emotion 'Surprise' is depicted in the Fig.5 given below. Similar files are obtained for each and every speech sample present in the database.

```
C:\htk\emorecog\data\train\mfcc>HList -e 1 -o -h su11.mfcc
------------------------------ Source: su11.mfcc ------------------------------
  Sample Bytes:  78        Sample Kind:   MFCC_D_A_C_K_0
  Num Comps:     39        Sample Period: 40000.0 us
  Num Samples:   105       File Format:   HTK
---------------------------- Observation Structure ----------------------------
x:      MFCC-1   MFCC-2   MFCC-3   MFCC-4   MFCC-5   MFCC-6   MFCC-7   MFCC-8   MFCC-9   MFCC-10
        MFCC-11  MFCC-12  C0       Del-1    Del-2    Del-3    Del-4    Del-5    Del-6    Del-7
        Del-8    Del-9    Del-10   Del-11   Del-12   DelC0    Acc-1    Acc-2    Acc-3    Acc-4
        Acc-5    Acc-6    Acc-7    Acc-8    Acc-9    Acc-10   Acc-11   Acc-12   AccC0
------------------------------ Samples: 0->1 ------------------------------
0:     -1.207    0.030    8.389    4.136   -4.815    9.456   -1.355    1.255    7.136   -0.800
        6.528    4.957   59.723   -0.016    0.124   -0.194    0.316    0.023    0.058    0.194
        0.424    0.104   -0.097   -0.364    0.162   -0.042    0.007   -0.013    0.046   -0.060
        0.060    0.044    0.006    0.026   -0.097   -0.031    0.053   -0.048    0.009
1:     -1.168    0.059    7.580    4.723   -4.437    9.782   -1.761    2.651    7.741
```

Fig. 5: MFCCs in HTK format

C. *Recognizer:*

Once the training phase is over, the recognizer comes into role. Its task is to assign an emotional label to the query used for testing the intelligence of HMMs. HMMs gives the output in the form of log probability with respect to all the classes present and assigns a label to the class having highest log probability. The emotional classes are shown in Fig. 6.

Fig. 6: Emotional Classes

III. DATABASE

The standard SAAVE database has be taken for the emotional classification. The four classes: sad, disgust, fear and surprise are chosen for this research. The utterances of the selected database were spoken by four male actors in the British English accent. 50 audio clips are used for training and 30 utterances are used for testing. The sentences used by this database have been picked from TIMIT database and justifies the labelled emotions. The training and testing database are non-overlapping which makes our system speaker independent. Some of the sample sentences from the SAAVE database are shown in table 1.

Table 1: Sample Audio clips from SAAVE

She had your dark suit in greasy wash water all year.
Don't ask me to carry an oily rag like that.
Will you tell me why?
Destroy every file related to my audits.
A few years later the dome fell in.

IV. RESULTS AND DISCUSSION

The system was trained with 50 emotional continuous speech samples from 4 different emotional classes. The performance of this emotion recognition system was calculated for an unseen set of data. The testing was done on 30 speech samples. The HVite tool of the HTK Toolkit was used to test the unseen data from the emotional corpus. HVite used the Viterbi decoding algorithm to do the classification. During the training and testing, the audio clips are segmented first and the silence part was not considered. Some parameters like frame size, overlap, pre- emphasis etc. were also optimized to attain high accuracy. Each HMM is trained by the MFCCs taken from the words only. After performing several experiments and trials the average accuracy of 78% and highest accuracy of 91.25% was achieved. The results in HTK format are shown in Fig. 7.

```
HResults -A -D -T 1 -e ??? sil -I testref.mlf def\labellist.txt rec.mlf

No HTK Configuration Parameters Set

=================== HTK Results Analysis ===================
  Date: Fri May 15 11:05:51 2015
  Ref : testref.mlf
  Rec : rec.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=91.25 [H=73, S=7, N=80]
WORD: %Corr=91.25, Acc=91.25 [H=73, D=0, S=7, I=0, N=80]
============================================================

No HTK Configuration Parameters Set
```

Fig. 7: Results in HTK format

The formula used by HTK to calculate the accuracy is given by Eq. 3

$$\%\ Accuracy = \frac{N - D - S - 1}{N} \times 100 \ldots\ldots\ldots (3)$$

Here, N is the total number of data samples, S is the number of substitution errors, D is the number of deletion errors, and I is the number of Insertion errors and H is the number of correctly labeled emotion utterances. Based on these results the confusion matrix was drawn and shown in table II.

Table II: Confusion Matrix Generated

Recognized Emotion:	Original Emotions			
	Surprise	**Sadness**	**Fear**	**Disgust**
Surprise	27	0	0	0
Sadness	1	30	1	3
Fear	2	0	29	0
Disgust	0	0	0	27

V. CONCLUSION AND FUTURE WORK

In this research paper, four emotional states were explored that is surprise, sadness, fear and disgust. The optimization of the HMM model was performed by ignoring the silence part present in the continuous speech and modifying the pre-

processing of the acoustical parameters. Thus the average accuracy of 78% and the best accuracy of 91.25% was achieved. The results obtained are very encouraging.
In this project, only the speech signals were processed to recognize different emotions, but we plan to extend this work further on classification of more emotions like happy, neutral, angry etc. and integrate this system with facial emotions recognition system.

Acknowledgement

Research described in this paper was supported in parts by Department of Science and Technology, Govt. of India under grant No. "DST/TSG/ICT/2013/37" and by the National Sustainability Program under grant No. LO1401. International cooperation in frame of COST IC1206 was supported by Czech Ministry of Education under grant No. LD14091.

References

[1] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller, "Cooperative Learning and its Application to Emotion Recognition from Speech", 2015.

[2] Ashish B. Ingale, D. S. Chaudhari "Speech Emotion Recognition", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[3] J. Sirisha Devi et al., "Automatic Speech Emotion and Speaker Recognition based on Hybrid GMM and FFBNN", International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.1, February 2011.

[4] K Sreenivasa Rao, Ramu Reddy, Sudhamay Maity and Shashidhar G Koolagudi "Characterization of emotions using the dynamics of prosodic features" West Bengal, India.

[5] F. Metze, T. Polzehl, and M. Wagner, "Fusion of acoustic and linguistic speech features for emotion detection," in Proc. International Conference on semantic Computing (ICSC). Berkeley, CA; USA: IEEE, Sep. 2009.

[6] F. Metze, T. Polzehl, and M. Wagner, "Fusion of acoustic and linguistic speech features for emotion detection," in Proc. International Conference on Semantic Computing (ICSC). Berkeley, CA; USA: IEEE, Sep. 2009.

[7] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Relevace of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functional. Pages 2253{2256 of: ISCA, Proceedings Interspeech.

[8] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," Speech Communication, vol. 40, no. 1-2, pp. 33–60, 2003.

[9] D. Ververidis and C. Kotropoulos, "A review of emotional speech databaes," in PCI 2003, 9th Panhellenic Conference on Informatics, November 1-23, 2003, Thessaloniki, Greece, pp. 560–574, 2003.

[10] B.Schuller, G. Rigoll, M. Lang, "Hidden Markov model-based speech motio recognition", Proceedings of the IEEE ICASSP Conference on Acoustis, Speech and Signal Processing, vol.2, pp. 1-4, April 2003

[11] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li, "Mandarin emotional speech eognition based on SVM and NN", Proceedings of the 18th International Conference on Pattern Recognition (ICPR/06), vol. 1, pp. 1096- 1100, September 2006.

[12] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin ,Ruhina Quazi, "Comparison between k-nn and svm method for speech emotion recognition", International Journal on Computer Science and Engineering (IJCSE).

[13]Anurag Kumar, Parul Agarwal, Pranay Dighe1, "Speech Emotion Recognition by Ada Boost Algorithm and Feature Selection for Support Vector Machine".

14] Li T., Ogihara, M., "Detecting emotion in music", in the Proc. International ymposium on Music Information Retrieval, Washington D.C., USA, page(s): 157-163, 2003

[15] S. D. Shirbahadurkar, A. P. Meshram, AshwiniKohok&SmitaJadhav "An Overview and Preparation for Recognition of Emotion from Speech Signal with Multi Modal Fusion IEEE Proceedings, Vol.5., 2010.

[16] VibhaTiwari, "MFCC and its applications in Speaker Recognition", Published in International Journal on Emerging Technology, ISSN No: 0975-8364, page 19-23, April 2010.