Emotion Recognition from Hindi Speech using MFCC and Sparse DTW

Er. VipinKumar R. Pawar
Head of Reasearch & Analysis Department
Anita's MicroSystems(INDIA), Nashik, INDIA

Ms. Nupur Patel
B.E.(Electronics & Telecommunication Engineering)
St. Francis Institute of Technology, Mumbai

***Abstract-:* Recently increasing attention has been directed to the study of emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. The project of Emotion Recognition from Hindi Speech address to three main aspects of speech recognition system. The first one is the choice of suitable features for speech representation. Using Sparse DTW for feature recognition has improved space efficiency and time complexity. Implementation of automatic emotion recognition system (using MATLAB) provides an accuracy of over 75% for 5 emotions namely: happy, sad, surprise, anger and neutral over a database containing large variety of speakers.**

1. INTRODUCTION

Speech is a vocalized form of human communication. Emotions exert an incredibly powerful force on human behaviour. Emotion plays an important role in a person's approach to a particular situation at that particular time. Unable to understand a person's emotion in a particular situation may cause a failure of communication. Thus recognising the emotion becomes one of the important aspects. This project mainly aims to classify 5 emotions namely sad, happy, anger, surprise and neutral. The input signal is divided into various frames of 20ms and features are extracted from each frame using MFCC. Later on, Sparse DTW is used for classification of emotions.

2. LITERATURE SURVEY

An experimental study on vocal emotion expression and recognition and the development of a computer agent for emotion recognition. They used RELIEF-F algorithm for feature selection. The total average accuracy is about 70%.[1]

Speech emotion recognition is done by use of continuous hidden Markov models. Two methods are propagated and compared throughout the paper. Within the first method a global statistics framework of an utterance is classified by Gaussian mixture models. [2]. Mel Frequency Cepstrum Coefficient(MFCC) feature has been used for designing a text independent speaker identification system. The goal of this project was to create a speaker recognition system, and apply it to speech of an unknown speaker. It also suggests some modifications to the existing techniques of MFCC for feature extraction to improve the speaker recognition efficiency. [3]

3. EMOTION RECOGNITION

3.1 BLOCK DIAGRAM

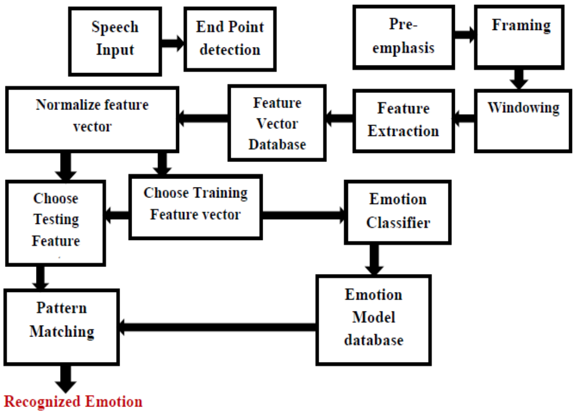


Fig. 1. Block diagram of the project

The speech signal is recorded for duration of 2sec and which has sampling frequency of 8000Hz. This is given as an input to end point detection algorithm (section 2.4) which removes the silence part. The output is then given to the SOLAFS for reducing the length of the frame but without affecting the pitch of the signal. The compressed signal is then given to pre-emphasis for boosting the high frequencies and framed into 20ms and then windowed to reduce the discontinuities. The MFCC features are then extracted and given to the classifiers (HMM, GMM) to evaluate the emotion. During the training period the output of classifier is stored into the emotion model database. During the testing period the classifier uses the speech database to identify the correct emotion.

3.2 SPEECH

Speech is the expression of thoughts and feelings using spoken language. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units. Humans use more than their ears while listening. They use the knowledge that they have about the speaker and the subject. There is a grammatical structure and redundancy that humans use to predict words not yet spoken. In computer Science, Speech recognition is translation of spoken words into text. It is also known as Automatic Speech Recognition (ASR). In ASR we only have the speech signal. It is difficult in ASR to accurately comprehend human

3.3 EMOTION

Emotion is a mental state that arises spontaneously rather than through conscious effort and is often accompanied by physiological changes, example: the emotion of joy, sorrow, hate, love etc. The word emotion includes wide range of observable behaviour, expressed feelings and change in the body state. Emotion is not only recognized through body language but also through speech. There is a huge list of human emotion that we are capable of experiencing. However often times we only experience very limited number of emotions for example happiness, love, stress, relax etc.

3.4 PRE-PROCESSING OF SPEECH

3.4.1 END POINT DETECTION

An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the endpoint detection problem.In this project End Point Detection Algorithm (EPD) is used for pre-processing of speech. An advantage of a good endpoint detecting algorithm is that proper allocation of regions of speech can substantially reduce the amount of processing required for the intended application.

This improves the performance of the decision making block and makes the system memory efficient because the templates produced by the feature extraction stage correspond to the detected speech only [2]. There are various methods for end point detection such as VAD (Voice activity detection) [3], algorithm on Mahalanobis Distance [3], algorithm on Energy Threshold [3]. For this example, the word “four”, it is not important to include the entire initial unvoiced interval; in fact, experience has shown that 30 ms to 50 ms of unvoiced energy is sufficient for most word recognition purposes [1].

4. FEATURE EXTRACTION AND RECOGNITION

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. In order to find some statistically relevant information from incoming data, it is important to have mechanisms for reducing the information of each segment in the audio signal into a relatively small number of parameters, or features. [4].

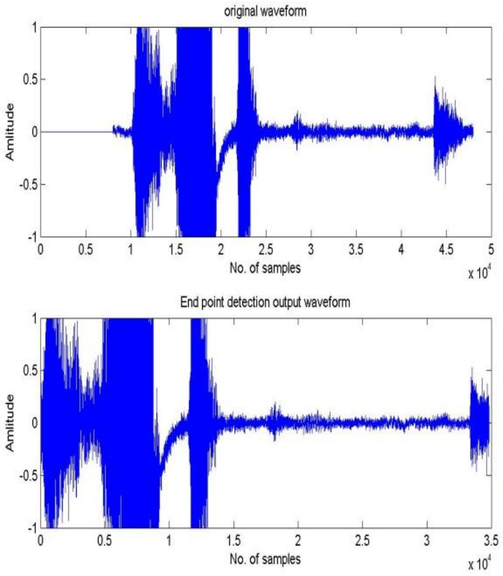


Fig. 2. End Point detection

Any emotion from the speaker’s speech is represented by the large number of parameters which is contained in the speech and the changes in these parameters will result in corresponding change in emotions. There are different features present in an emotion like pitch, energy, duration format, Mel Frequency Cepstrum Coefficient (MFCC) [5] and Linear Prediction Cepstrum Coefficient (LPCC) .When there is change in emotional state there is corresponding change in speech rate, energy, and spectrum.

4.1 MEL FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)

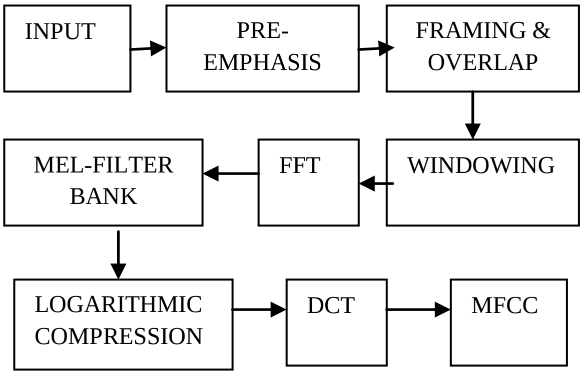


Fig. 3. MFCC block diagram

4.1.1 INPUT AND PRE-EMPHASIS

The pre-emphasis is a preprocessing phase which increases, within a band of frequencies, the magnitude of higher frequencies with respect to the magnitude of lower frequencies. In speech processing, the original signal usually has lower frequency energy, and processing the signal to emphasize higher frequency energy is necessary. The filter transfer function is given by,

$$\mathrm{H(z)} = 1 - \mathrm{az}^{-1}$$

Where, ‘a’ is between 0.9 and 1.

Pre-emphasis of a single frame after taking FFT is shown in figure 4 and figure below.

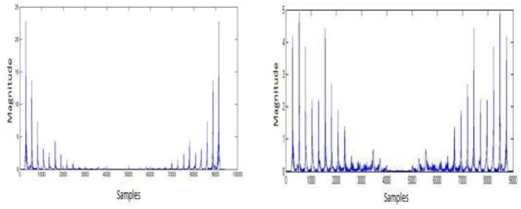


Fig.4: Before Pre-emphasis after Pre-emphasis

4.1.2 *FRAMING AND OVERLAPPING*

Normally a speech signal is not stationary, but seen from a short-time point of view it is. Typically, a speech signal is stationary in windows of 20 ms. Therefore the signal is divided into frames of 20 ms which corresponds to n samples:

$$n = t_{st} \times f_s$$

When the signal is framed it is necessary to consider how to treat the edges of the frame. Therefore it is expedient to use a window to tone down the edges. As a consequence the samples will not be assigned the same weight in the following computations and for this reason it is prudent to use an overlap.

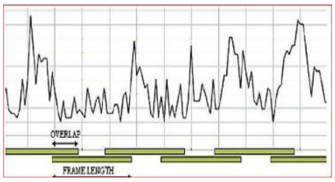


Fig. 6: Illustration of framing: The speech is divided into four frames

4.1.3 *WINDOWING*

Awindow functionis a mathematical function that is zero-valued outside of some chosen interval.In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each one i.e.the Hamming window. Hamming window is described as follows:

$$H(n)=0.54+0.46\cos(2n/(n-1))$$

From the above figure it can be seen that the amplitude of the side lobes is smaller than the amplitude of main lobe. Thus, windowing minimizes the spectral distortion and the discontinuties at the beginning aand end of each frame.

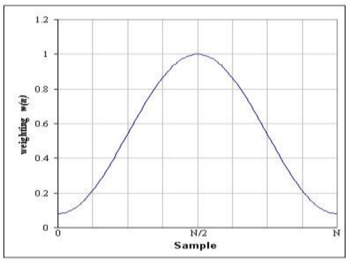


Fig. 7: Hamming window

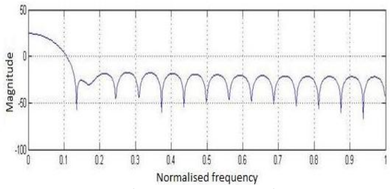


Fig. 8: Normalized frequency plot of hamming window

4.1.4 *MEL FILTER BANK*

The Mel filtering approximates the non-linear characteristics of the human auditory system in frequency. The output is an array of filtered values, typically called mel-spectrum, each corresponding to the result of filtering the input spectrum through an individual filter. Therefore, the length of the output array is equal to the number of filters created.

$$MelFrequency = 2595 * \log(1 + lf/700)$$

Where, lf= linear frequency.

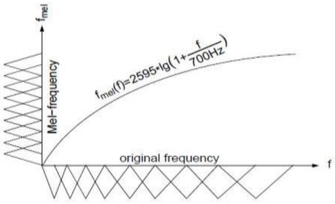


Fig. 4.7 Mel filter bank

4.1.5 *DISCRETE COSINE TRANSFORM*

DCT[6] is applied on the log Mel filter bank output coefficients to get Mel-scale Cepstral Coefficient. DCT can be given as :

$$c(n) = \sum_{m=0}^{M-1} s(m) \times \cos\left(\frac{\pi n(m-0.5)}{m}\right)$$

Where, n=1,2,3, ..,L
M = number of triangular band pass filters.
Here, M= 33 and L= 12.

4.1.6 *LOG ENERGY*

Log energy can be obtained from a signal to increase the accuracy of the system as it is an important feature. Thus we add log energy as the 13th coefficient of MFCC. Log energy can be obtained by using following equation:

$$\log E = \log \sum_{n=1}^{N} (x(n)^2)$$

4.1.7 *LIFTERING*

Liftering is a signal processing technique in which undesirable spectral measurement variations can be partially controlled (i.e., reduced in the level of variation). In particular, a band pass liftering process [7] reduces the variability of the statistical components of MFCC-based spectral measurements and hence it is desirable to use such a liftering process in a speech recognizer.

4.1.8 DELTA MFCC COEFFICIENT

Delta is the time derivative of any signal. We perform delta function i.e. time derivative of the output from the previous stage (energy +MFCC coefficients). Thus it gives us the velocity and acceleration out of the obtained 13 coefficients. Delta provides us with 26 coefficients (13 normal and 13 derivatives). Delta MFCC can be obtained by using equation:

$$\Delta c(n) = \left(\sum_{i=1}^{D} i.\,(c(n+i) - c(n-i))\right) \Big/ \left(\sum_{i=1}^{D} i^2\right)$$

4.1.9 DOUBLE DELTA MFCC COEFFICIENT

Double Delta MFCC coefficients are the time derivatives of the Delta MFCC coefficients. Double Delta provides us with 39 coefficients (13 normal, 13 derivatives amd 13 double derivatives).

5. FEATURE RECOGNITION

5.1 SPARSE DTW

The main principle of using Sparse DTW is to reduce time and space complexity. In order to reduce space usage while avoiding any recomputations we consider the following facts:

1. Quantizing the input time series to exploit the similarity between the points in the two series.
2. Using a sparse matrix of size k,
   In the worst case. K = m×n
   If the two sequences are similar k << n× m.
3. The warping matrix is calculated using dynamic programming and sparse matrix indexing.

5.1.1 SPARSE APPROACH

Let us consider two sequences:
S = [3, 4, 5, 3, 3] and Q = [1, 2, 2, 1, 0].
First quantize the sequences into the range [0 - 1] using Equation

$$\text{Quantized Seq}_i^k = \frac{s_i^k - \min(s^k)}{\max(s^k) - \min(s^k)}$$

Where, s_i^k denotes the ith element of the kth time series. This yields the following sequences:

S' = [0, 0.5, 1.0, 0.0, 0.0] and Q' = [0.5, 1.0, 1.0, 0.5, 0.0]

Next create overlapping bins governed by two parameters: bin-width and the overlapping width. For this particular example, the bin-width is 0.5. Thus 4 bins obtained are shown in table below:

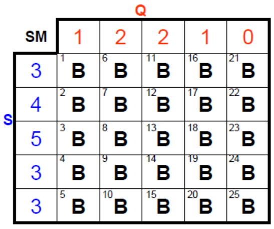
Bin Number (B_k)	Bin Bounds	Indices of S'	Indices of Q'
1	0.0-0.5	1,2,4,5	1,4,5
2	0.25-0.75	2	1,4
3	0.5-1.0	2,3	1,2,3,4
4	0.75-1.25	3	2,3

Table: Bin bounds where B_k is the kth bin.

. It can be represented in much less than n × m space, where n and m are the lengths of the time series S and Q, respectively.

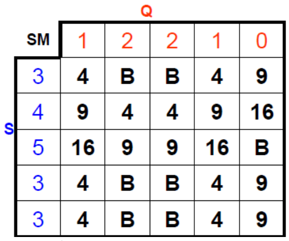
$$SM(i,j) = \begin{cases} \text{EucDist}(S(i), Q(j)) & \text{if } S(i) \text{ and } Q(j) \in B_k \\ B & \text{otherwise} \end{cases}$$

Unblocking (opening) the cells that reflect the similarity between points in both sequences, the SM entries are shown in Figure 9.



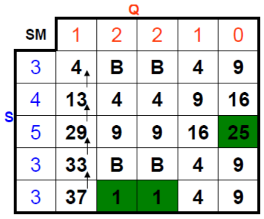
SM	1	2	2	1	0
3	1 B	6 B	11 B	16 B	21 B
4	2 B	7 B	12 B	17 B	22 B
5	3 B	8 B	13 B	18 B	23 B
3	4 B	9 B	14 B	19 B	24 B
3	5 B	10 B	15 B	20 B	25 B

Fig. 8: SM initial blocked[B]



SM	1	2	2	1	0
3	4	B	B	4	9
4	9	4	4	9	16
5	16	9	9	16	B
3	4	B	B	4	9
3	4	B	B	4	9

Fig. 9: SM after unblocking the optimal cells.

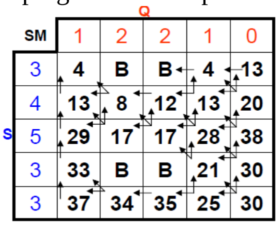


SM	1	2	2	1	0
3	4	B	B	4	9
4	13	4	4	9	16
5	29	9	9	16	25
3	33	B	B	4	9
3	37	1	1	4	9

Fig. 10: Unblocking upper neighbor (Shaded cell).

Then calculate the warping cost for each open cell c ∈ SM (cell c is the number from the linear order of SM's cells) by finding the minimum of the costs of its lower neighbors, which are [c - 1; c - n; c - (n+1)] (black arrows in Figure 11 show the lower neighbors of every open cell). This cost is then added to the local distance of cell c. The above step is similar to DTW, however, we may have to open new cells if the upper neighbors at a given local cell c∈ SM are blocked. The indices of the upper neighbors are [c + 1; c + n; c + n + 1], where n is the length of sequence S (i.e., number of rows in SM).If the Upper Neighbors = 0 for a particular cell, its upper neighbors will be unblocked. This is very useful when the algorithm traverses SM in reverse to find the final optimal path. In other words, unblocking allows the path to be connected. For example, the cell SM (5) has one upper neighbor that is cell SM (10) which is blocked (Figure 9), therefore this cell will be unblocked by calculating the EucDist(S (5), Q (2)). The value will be added to the SM which means that cell SM (10) is now an entry in SM (Figure 10). Although unblocking ads cells to SM which means the number of open cells will increase, but the overlapping in the bins boundaries allows the SM's unblocked cells to be connected mostly that means less number of unblocking operations.

Figure 11 shows the final entries of the SM after calculating the warping cost of all open cells.

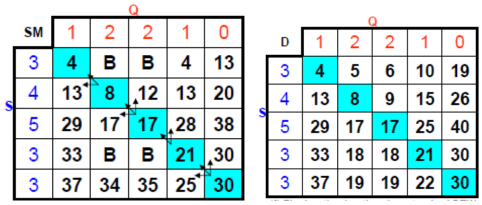


SM	1	2	2	1	0
3	4	B	B	4	13
4	13	8	12	13	20
5	29	17	17	28	38
3	33	B	B	21	30
3	37	34	35	25	30

Fig. 11: Constructing SM.

Hop initially represents the linear index for the (m, n) entry of SM that is the bottom right corner of SM in Figure 12. Starting from hop = n×m choose the neighbors [hop–n, hop–1, hop - (n+1)] with minimum warping cost and proceed recursively until the first entry of SM is reached, namely SM (1) or hop = 1. While calculating the warping path only look at the open cells, which may be fewer in number than 3. The filled cells show the optimal warping path, which crosses the grid from the top left corner to the bottom right corner. The distance between the two time series is calculated using Equation:

$$DTW(S,Q) = \left\{ \frac{\sqrt{\sum_{k=1}^{K} W_k}}{K} \right.$$



SM	1	2	2	1	0
3	4	B	B	4	13
4	13	8	12	13	20
5	29	17	17	28	38
3	33	B	B	21	30
3	37	34	35	25	30

(I)

D	1	2	2	1	0
3	4	5	6	10	19
4	13	8	9	15	26
5	29	17	17	25	40
3	33	18	18	21	30
3	37	19	19	22	30

(II)

Fig. 12: Final optimal path using (I)SparseDTW. (II)DTW.

6. DATABASE DESCRIPTION

The database includes speech samples from around 50 people which were collected manually for five different emotions namely: Sad, Happy, Anger, Neutral and Surprise. These samples were collected using a software named Audacity and also by writing a program in MATLAB using Data Acquisition Toolbox. Phrases are recorded using a microphone.

Various sentences used for collecting database are as follows:

- *Happy Emotion:* Ajj me bhot khush hu.
- *Sad emotion:* Mujhe koi yaad hi nahi karta.
- *Neutral Emotion:* Me ghar ja raha hu.
- *Anger Emotion:* Mujhe gussa maat dilao.
- *Surprise Emotion:* Kya baat kar rahe ho!

7. CONCLUSION

This implementation of automatic emotion recognition system(using matlab) provides an accuracy of over 75% for 5 emotions namely: happy, sad, surprise, anger and neutral. In this project Mel Frequency Cepstral Coefficients (MFCC) is used for feature extraction and Sparse DTW is used for feature recognition purpose.Using SparseDTW helped to improve the space efficiency and reduce the time complexity. Progress in the area relies heavily on the development of appropriate databases. The problems that usually occur while collecting database are variations in the surrounding(noise), varient speaker characteristics and acoustic confusability.

8. FUTURE SCOPE

This project is an initiative to identify, explore and develop possible alternatives to improve the overall human-computer interaction. The scope for this project includes using HMM for further increasing the accuracies and reducing time complexity. Also the database can be made multilingual(including different languages together). Further this project can be converted into an android application.

9. REFERENCES

(1) L. R. Rabiner, M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal*, 54, p. 297-315, Feb. 1975.

(2) T. B. Amin, I. Mahmood, "Speech Recognition Using Dynamic Time Warping", *2nd International Conference on Advances in Space Technologies, Proceedings of ICAST*, vol. 2, pp. 74-79, November, 2008.

(3) K. Yamamoto, F. Jabloun, K. Reinhard and A. Kawamura, "Robust method for end point detection using discriminative feature extraction", *IEEE Proceedings, Europe*, 2006.

(4) K. R. Aida–Zade, C. Ardil and S.S. Rustamov , " Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems ", *World Academy of Science, Engineering and Technology*, 2006.

(5) Digital Signal Processing Mini-Project, "An Automatic Speaker Recognition System", *Minh N. Do, Audio Visual Communications Laboratory*, Swiss Federal Institute of Technology, Lausanne, Switzerland.

(6) L. Rabiner, and B. Juan, Fundamentals of speech recognition, Prentice Hall PTR, New Yersey, ISBN 0-13-015157-2.

(7) F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech", *IEEE International Conference on Emotion and Signal Processing*, pp. 1970-1973, 2004.