ISSN:0975-9646

Speech Database for Speech Analysis

Ganga Banavath[1], Sreedhar Potla[2]

[1]*M.Tech Student, IT Department, Sreenidhi Institute of Science and Technology, India*

[2]*Associate Professor, IT Department, Sreenidhi Institute of Science and Technology, India*

Abstract: Speech processing is becoming an emerging technology which enables interaction of human beings with machines. We are considering a standard length of speech sample which gives higher performance in speech processing. Our aim is to describe an efficient technique or method resulting effective speech processing applications. Performance of a classifier is a function depends on the length of speech sample, environment etc. This work is carried out using Mel frequency cepstral coefficients (MFCC) along with Gaussian Mixture Model (GMM) classifier. Results are evaluated on our own database comprising five South Indian Languages. The higher recognition performance is observed to be 82.6531% for 10seconds speech, 60.2041% for 5seconds speech and 57.1341% for 3seconds speech.

Keywords: Speech Database, Mel-frequency cepstral coefficients (MFCC), Gaussian Mixture Model (GMM).

I. Introduction

Speech is a natural means of communication for humans. Human beings can recognize the speech, speaker, language and utterance by hearing the speech. About 2-3 seconds of speech is sufficient for a human to identify a speech, speaker, language and utterance. One review on human speech recognition states that many studies of 8-10 speakers yield accuracy of more than 97% if a sentence or more of the speech is heard. Performance falls if the length of the speech is short and if the numbers of speakers are more. We considered different lengths of speech samples (3seconds, 5 seconds and 10 seconds) as input for speech analysis. We can classify the speech in two ways: 1. Text-dependent vs. Text-independent speech, 2. Noisy speech vs. noiseless speech. Here we are using Text-independent noiseless speech for speech analysis.

Till today many researchers developed a variety of speech databases for analyzing the performance of speech applications, but they lack the efficiency to build the effective speech applications due to less duration (in terms of time) of speech data available on them. Here we are providing the details of one of the widely used speech databases in speech analysis is TIMIT database, it is the first largest database consists recordings of 630 speakers, each speaker reading 10 sentences with typical sentence duration of 5 seconds only. This is insufficient for many speech analysis applications [2]. For this reason we manually collected speech samples and prepared our own speech database and details of database are provided in the following section. On this database we are performing speech analysis. The speaker specific acoustic features of the speech can be retrieved by using Mel-frequency cepstral coefficients (MFCC) algorithm [4][5] which represents efficient feature vector. To evaluate the performance of speech analysis system, Gaussian mixture model (GMM) classifier is used along with Expectation Maximization (EM) algorithm [1][9] .

Rest of the paper is organized as follows. The details of proposed speech database are discussed in Section II. Section III, discusses about system description. Section IV discusses about building the model. Section V discusses about Observed Results. Conclusion has been discussed in the final section.

II. Speech Database

Speech samples of the proposed database have been collected from Prasar Bharathi [4] for five different languages namely Hindi, Indian English, Tamil, Telugu and Kannada. The speech corpus is collected from news bulletins and is considered for speech analysis. From each speaker, about 5 to 10 minutes of speech is collected. On the whole, each language contains minimum of 2.7 hours of speech data. Details of the database are described in Table I. The main reasons for choosing Radio channel for collecting the proposed database are: It is difficult to find a sufficient number of native speakers in a single place, It is extremely time consuming to record the required speech data from different speakers at different places and the speakers from radio channels are professional and matured enough at pronunciation and speaking styles. The speech considered contains a problem that the news bulletins contain background music during the headlines at the start and end of the speech. Speech database preparation involves recording and collecting the speech samples and storing them into a storage database. Therefore, while preparing the database proper care has been taken to consider only the speech content without background music in it. Figure Fig. 1 represents the process of speech database preparation.

From source Prasar Bharathi [4] we captured the speech files and stored them in a storage data. After that we prepared them according to the language and stored them in “Language” storage data and to name the languages we followed a standardized naming convention (L1, L2, L3, L4, L5). In each language we have a different number of female and male speakers, therefore we followed a standardized naming conventions (Speaker 1,..., Speaker n) and we stored the speech data according to the respective speakers. Here train wave files storage data represents the speech samples used for training purpose and test wave files storage data represents the speech samples used for testing purpose.

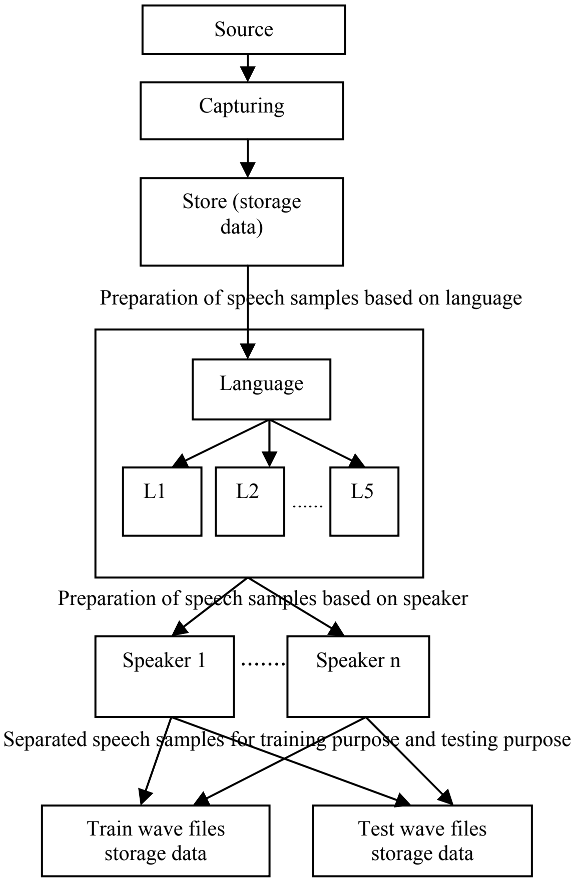
Source
Capturing
Store (storage data)
Preparation of speech samples based on language
Language
L1
L2
......
L5
Preparation of speech samples based on speaker
Speaker 1
....
Speaker n
Separated speech samples for training purpose and testing purpose
Train wave files storage data
Test wave files storage data

Fig. 1 Representation of speech preparation

In our speech database we have 15.37 hours of speech data for all the five languages consisting of 110 different speakers. For training we considered 80% of speech and for testing we considered 20% speech from each speaker respectively for **speech analysis**. In the below table we are providing the details of the speech data.

TABLE II
DETAILS OF COLLECTED SPEECH SAMPLES

S.No.	Language	No of speakers		Duration of speech considered (in minutes)
		Female	Male	
1	Indian-English	16	11	204.50
2	Hindi	15	12	196.35
3	Tamil	10	07	170.08
4	Telugu	08	12	174.92
5	Kannada	10	09	176.91

III. SYSTEM DESCRIPTION

Any speech analysis process mainly involves following steps: Speech database preparation, Pre-processing and Segmentation, Feature Extraction using MFCC and Building the Model using GMM with combination of EM. Speech database preparation has been discussed earlier in Section III. Here we are providing the brief introduction about pre-processing and segmentation and feature extraction. Building the model will be described in the next section.

A. Pre-processing and Segmentation

In this phase we are going to segment the speech samples into the length of 3seconds, 5seconds and 10seconds that which contains only the speech but not the noise. So to remove the music and to segment speech samples into smaller ones we had done segmentation. As we mentioned there is music at the start and end of speech samples we neglected the music and considered only speech portion and stored it into a storage data. To do so, we captured the speech sample and read the speech data consisting of 3, 5 and 10 seconds and stored it in a storage data by providing a standardized naming convention as per our convenience. We repeated this process until we reach the end of the speech sample as well as we repeated this process for all the speech samples. Figure Fig. 2 represents the process of segmentation.

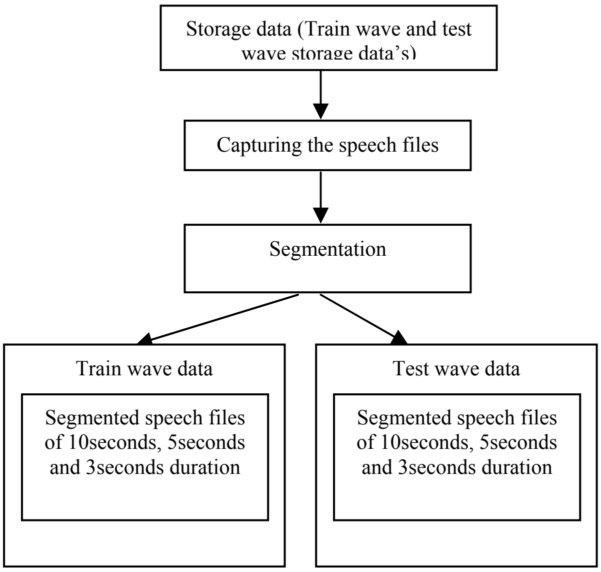
Storage data (Train wave and test wave storage data's)
Capturing the speech files
Segmentation
Train wave data
Segmented speech files of 10seconds, 5seconds and 3seconds duration
Test wave data
Segmented speech files of 10seconds, 5seconds and 3seconds duration

Fig. 2 process of segmentation

In Table IV we are providing the details of speech corpus after segmentation of speech data.

TABLE V
DETAILS OF SEGMENTED SPEECH SAMPLES

Language	No of speakers		Length of speech considered (in minutes)	No of segmented samples collected	
	Female	Male		female	Male
English	16	11	204.50	789	351
Hindi	15	12	196.35	571	554
Tamil	10	07	170.08	704	400
Telugu	08	12	174.92	434	639
Kannada	10	09	176.91	597	455

B. Feature Extraction

Feature extraction is the process of transforming the speech signal in to a set of feature vectors. The feature vector represents the **speaker specific** information due to vocal tract, excitation source and behavioural traits. A good feature vector set should have representation of all the components of speaker information. MFCCs are mostly related to the human peripheral auditory system. The main purpose of the MFCC is to mimic the behaviour of the human ears [7]. According to the studies, human hearing does not follow the linear scale but rather then it follows Mel-spectrum scale which is a linear spacing at low frequencies below 1 KHz and logarithmic scaling at high frequencies above 1 KHz to capture the important characteristics of the speech. The figure Fig3 represents the block diagram of MFCC and also we are providing brief of mfcc steps.

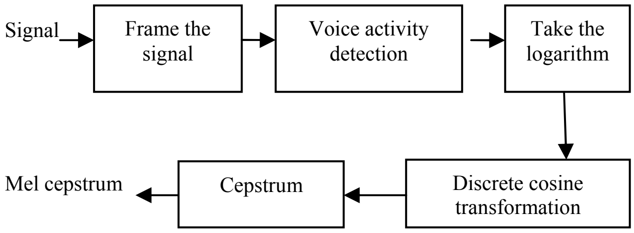

Fig. 3 Block diagram of MFCC process

1) *Frame the Signal into Short Frames*: We can observe that a speech signal is constantly changing, so to simplify the things we are representing the speech signal on 25 milliseconds window and we considered 10 milliseconds hop time (steps between successive windows) of each window.
2) *Voice Activity Detection*: To do so we are applying Discrete Fourier Transformations of the frames to find out power spectrum of frames and after that we need to compute filter bank energies of each frame.

   $S_i(k) = \sum_{n=1}^{N} S_i(n)h(n)e^{-\frac{j2\pi kn}{N}}$, $1 \le k \le K$-- (1)

   Where $S_i(k)$ is the time domain frame, S(n) is time domain signal, i is frame number, h(n) is hamming window of sample length N, K is the length of DFT.

   $P_i(k) = \frac{1}{N}|S_i(k)|^2$----------------------------- (2)

   Where $P_i(k)$ is power spectrum of frame i.
   Formulae for computing mel-filter bank energies:

   $mel(f) = 2595 * log10(1 + \frac{f}{700})$----------- (3)

   Here f is frequency and which represents $P_i(k) \forall\, i, k$.
3) *Take the Logarithm of All Filter Bank Energies*: Once we have the filterbank energies, we need to take the logarithm of them. This is because we don't hear loudness on a linear scale, to do so we need to add higher energy approximately 8 times energy to it. This compression operation makes our features match more closely what humans actually hear. Why we apply this logarithm is because it allows us to use cepstral mean subtraction, which is a channel normalisation technique.
4) *Discrete Cosine Transformations (DCT)*: of all the filterbank energies. Keep the DCT coefficients of 2-13 and discard the rest. This is because if we have more DCT coefficients then we will get fast changes in the filterbank energies and it degrades the performance, so to get a small improvement we need to drop them. Here we are providing the formulae for converting the frequencies into mel-frequencies.

IV. BUILDING THE MODEL

In this work, GMMs with combination of EM are used for developing the speech analysis systems using acoustic features. GMM is well known as a classifier, which is used to build the speech analysis model by capturing the distribution of data in feature space. The accuracy in capturing the true distribution of data depends on various parameters such as dimension of feature vectors, number of feature vectors and number of mixture components. In this work, GMMs are assumed to capture the speaker specific information from the given acoustic features. In this work, we have used Mel Frequency Cepstral Coefficients (MFCC) for extracting the features. 13 MFCC features are derived from a speech frame of 25 ms with a frame shift of 10 ms. for deriving the MFCCs, 24 filter bands are used.

For developing the speech analysis system using GMMs, we need to develop a specific GMM for each of the language. These GMMs are developed using the spectral vectors derived from the speech corresponding to the languages considered. To do so, EM algorithm is used to estimate the parameters from a given distribution of data. In this work, 5 Indian languages are considered for analyzing the speech performance using acoustic features. Therefore, the proposed speech analysis system consists of 5 GMMs (language models) developed using speech corresponding to 5 languages. For evaluating the developed speech analysis system, feature vectors derived from test speech samples are given to all the language models.

A. Gaussian Mixture Model

In general, a mixture model is a probabilistic model which assumes the underlying data is belongs to a mixture distribution. A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities that measures continuous features in various speech systems [8]. The parameters of Gaussian mixture model are mean (μ), co-variance matrix (∑) and it's probability density function is given by:

$$p(x) = \frac{1}{(2)^{\frac{k}{2}}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)\text{-------}(4)$$

Where p(x) is probability density of x and x is a data vector. The advantages of the GMMs include flexibility, general-purposeness and the existence of an effective estimation algorithm. There are several techniques available for estimating the parameters of a GMM [9] but most often GMMs are estimated using the Expectation-Maximization (EM) algorithm [10] for its maximum convergence estimation.

B. Expectation Maximization Algorithm

Expectation Maximization algorithm [11] [12] is an iterative algorithm that starts from some initial estimate of parameter or data set (e.g., random) and then proceeds to iteratively update the data set until convergence or maximum likelihood observation is detected. EM algorithm alternates between two phases namely E-step (expectation step) and an M-step (maximization step). In particular, EM algorithm attempts to find the parameters Θ that maximize the log probability $\log p(x;\ \Theta)$ of the observed data. The EM algorithm first finds the expected value of the complete data log-likelihood $\mathrm{logp}(\mathrm{x}, \mathrm{y}/\ \Theta)$ with respect to the unknown data Y given the observed data X and the current parameter estimates until the convergence reaches. That is, we define as follows:

$$\mathrm{Z}\big(\ \Theta\ ,\ \Theta^{\ (\mathrm{i}-1)}\big) = \mathrm{E}[\mathrm{logp}(\mathrm{x},\mathrm{y}|\ \Theta\ )|\mathrm{x}, \Theta^{\ (\mathrm{i}-1)}]\text{-----}\ (5)$$

Where, $\Theta^{(i-1)}$ are the current parameters estimates that we used to evaluate the expectation and Θ are the new parameters that we optimize to increase dataset (Z).

Any GMM application deals with 2 phases, training the model and testing the model. In training phase, feature vector creation will be done by using the extracted features of the speakers who are considered for training. In which we are applying 75 iterations to find out the maximum likelihood parameter estimation. This process is used to form the mixtures (here mixtures are nothing but clusters) of the data parameters. In testing phase, extracted features of the speakers who are considered for testing will be compared with the feature vectors which are created in previous phase (GMM training phase).

V. Observations

In this work, about 80% of the data is used for training the GMMs (developing the language models), and the rest 20% (which is not used for developing the models) data is used for testing or evaluating the performance of developed speech analysis models. Initially, we have developed speech analysis systems separately, using MFCC features. In this work, speech from all the speakers of a given language is used for developing and evaluating the models. We are considering three different lengths (3, 5 and 10 seconds) of test speech utterances for analysis. The performance of different lengths of test speech utterances is given in Table VI . By observing the results, the performance seems to be better with 10 seconds of speech compared to the both 5 seconds and 3seconds of the speech. The average performance for 10seconds wave file is 72.65306% and for 5sec wave file is 53.67348%.

Table VII
Final results of speaker analysis

Langue	No of speakers	Percentage match for 10sec speech input	Percentage match for 5sec speech input	Percentage match for 3sec speech input
Telugu	20	82.6531	47.9592	39.3592
Hindi	27	78.5714	60.2041	57.1341
English	29	69.3878	58.1633	56.1564
Kannada	19	64.2857	45.9184	41.8634
Tamil	17	68.3673	56.1224	53.2567

VI. Conclusion

This work has dealt with 5 different South Indian Languages and tested the usage of different length MFCC feature set with GMM (of different clusters 16 mixtures, 32 mixtures and 64 mixtures). The average performance has shown for 3 seconds speech is 49.55396%, 5 seconds speech is 53.67348% and for 10 seconds speech is 72.65306%. The efficiency of the system may be improved with the help of more number of mixtures and GMMUBM may be used for improving the performance. In future the work can be expanded to more number of languages.

References

[1] André Gustavo Adami, Automatic Speech Recognition: From the Beginning to the Portuguese Language, Universidade de Caxias do Sul, Centro de Computação e Tecnologia da Informação.

[2] Rua Francisco Getúlio Vargas, 1130, Caxias do Sul, RS 95070-560, Brasil.

[3] Stephen A. Zahorian, Jiang Wu, Montri Kamjanadecha, Chandra Sekhar Vootkuri, Brian Wong, Andrew Hwang, Eldar Tokhtamyshev: Open Source Multi-Language Audio Database for Spoken Language Processing Applications. INTERSPEECH 2011: 1493-1496

[4] http://www.newsonair.nic.in/.

[5] Urmila Shrawankar Research Student, (Computer Science & Engg.), SGB Amravati University; Dr. Vilas Thakare Professor & Head, PG Dept. of Computer Science, SGB Amravati University, Amravati TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH RECOGNITION SYSTEM: A COMPARATIVE STUDY.

[6] M. A. Anusuya, S. K. Katti, Speech recognition by Machine: A Review, (IJCSIS) International Journal of Computer Science and Information Security, 2009.

[7] Douglas Reynolds (MIT Lincoln Laboratory) “Gaussian Mixture Models”

[8] Jeff A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”, International computer science institute (ICSI).

[9] D. Reynolds, R.Rose, “Robust text-independent speaker identification using Gaussian Mixture Models”, IEEE Trans. Speech Audio Processing, VOL. 3, NO. 1, JANUARY 1995.

[10] D. Reynolds, “Gaussian Mixture Models*”, MIT Lincoln Laboratory,244 wood St. Lexinton, MA 02140,USA.

[11] D. A. Reynolds, “An overview of Automatic Speaker Recognition Technology”, MIT Lincoln Laboratory,244 wood St. Lexinton, MA 02140,USA,IEEE 2002.

[12] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38.

[13] http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm