CHAPTER V

SPEAKER VERIFICATION SUBJECT TO SPARSE SPEECH SIGNAL

5.1. INTRODUCTION

In spite of numerous contrasts in the middle of people, and the presence of numerous dialects, discourse takes after general examples, and all things considered has all around characterized attributes, for example, those of volume, recurrence dispersion, pitch rate and syllabic rate. These qualities have adjusted concerning environment, listening to and voice creation restrictions – discourse attributes fit the discourse producing capacities of the body – yet the quick changes in the public arena over the previous century have surpassed our capacity to adapt. The shouting mechanism for ‘long distance’ communications, for example, across an open valley, is not particularly suited to inner-city conditions (just stand outside a tower block for a few minutes on a hot day when windows are open and hear the examples of inappropriately loud vocal communications). On the other hand, rail or bus commuters will seldom have the opportunity to converse in whispers. The International Phonetic Alphabet (IPA) is the usual method of describing and writing the various phonemes that make up speech. As shown in figure 5.1. As defined by the International Phonetic Association, a set of symbols, written using a shorthand notation, describes the basic sound units of words.
These symbols can completely describe many different languages using the 107 letters and several diacritical marks available. It is beyond the scope of this book to introduce this alphabet, but simply to point out that researcher working with phonetics would be advised to learn the IPA and apply this notation in their work to avoid misconceptions and insufficiently specified speech sounds.

5.1.1. SPEECH CLASSIFICATION

Physically, the hints of discourse can be depicted as far as a pitch shape and formant frequencies. In fact this description forms a method of analysis used by most speech compression algorithms (discussed in Section 5.2 and beyond). Formants are thunderous frequencies of the vocal tract which show up in the discourse range as clear crests. As an example, three distinct formant peaks can be seen in the frequency domain plot of a short speech recording, in figure 5.2.
Figure 5.2: Spectrum plot of a 20 ms recording of voiced speech, showing three distinct formant peaks

The pitch contour (often called f0 – note the lower case notation) is the parameter that portrays the voice's tone (the apparent recurrence), and is basically the major vocal recurrence. Again, pitch frequencies contain energy but contribute little to intelligibility for English and other European languages. It is, however, a very different matter in a tonal language such as Mandarin Chinese which is totally dependent on tone for conveying meaning. As an example, in Chinese the single word ‘ma’ can mean one of five things depending on which tone it is spoken with: mother, horse, scold, question, etc. and this is not an isolated example since all single Chinese word sounds have multiple meanings differentiated by tone.

Formants have been portrayed by the well known analyst Klatt and others similar to the absolute most critical element in discourse interchanges. For the most part numerous formants will be available in a common articulation [85], and the area of these will shift after some time as the mouth's state changes. Formants are checked from the least recurrence upwards, and generally just the initial three (F1, F2 and F3) contribute essentially to the coherence of discourse. Some fricative
sounds like/ch/can creates a ton of formants, yet as a rule F1 contains the discourse’s majority vitality while F2 and F3 between them contribute more to discourse understandability.

In our research work we have considered a typical case where the speech data for training as well as for testing is very sparse. In some cases due to external circumstances like ill health, emotional state of the customer, environment in which voice is recorded; telephone communication where there will be lots of disturbances etc, and we get a sparse data after filtering the original speech signal. Here we can consider speech signals of very less duration like 2 ~ 6 sec for training as well as testing. Our objective is speaker verification using this sparse speech signals. But a sparse signal will not be sufficient for verification, so we have to fill these holes. This can be accomplished by getting acoustic information from these sparse signals and using mapping methods we can map the cohort speaker from a benchmark dataset. For mapping we used MAP adaptation and the benchmark dataset we used here is TIMIT corpus. Using the phonemes from the cohort speech signal we can fill the holes present in the sparse speech signal. To achieve this task we have come up with a new procedure where we index the phonemes of speech signals in an indexing table.

As the point of concentration is phonemes, the following section gives a brief description of their distribution. Then a detailed note on the proposed system is narrated where the importance of the proposed algorithm for feature selection is embedded to reduce the overhead of the problems raised by sparse speech signal.

### 5.2. DISTRIBUTION OF PHONEMES

A *phone* is a consonant or vowel speech sound. A *phoneme* is any equivalent set of phones which leaves a word. Three common associations of phonemes are shown in the following figure 5.3.
Figure 5.3: Phonemes represented in three different forms

Figure 5.3 (a) explains the first association among the phonemes, that all other class of phonemes are placed only after the accurate one. Next figure 5.3 (b) explains the second association, phonemes are placed in a layer like structure where the bottom layer consists of other phonemes and as we traverse towards the accurate root phoneme we will find more related phonemes. In the example we have seen that four layers of phonemes are placed in such a way that when we travel from leaf node to root, an accurate phoneme /oy/ is attained. The siblings of the same layer do not have any order preference. It shows that each layer consists of a subset of phonemes which are disjoint to each other. The arrangement of these subsets in the form of layers is based on their preferences. The final figure 5.3 (c) explains the third association among the phonemes in the form of a tree. From the root all the nodes except the leaves stand for a subset of class of phonemes. Only the leaves accommodate the phonemes. The relation between a parent node and a child node depicts the association among them. This illustration says that some kind of mistakes is more preferable than others.

One or more agglomerated phonemes can make up a syllable sound. For humans, it seems that the simplest unit of recognition may be the syllable, whereas the phoneme is a distinction generally
made by speech researchers aiming to determine a set of basic building blocks for speech sounds. Phoneme duration varies from language-to-language, speaker-to-speaker and differs depending upon the exact phoneme. There is even evidence to tie phoneme length to word stress – louder, and more emphasized phonemes tend to exhibit a longer duration. Division of phonemes in an informal voice of 10 minutes and a fake voice of 1 minute is shown in figure 5.4.

Figure 5.4: Division of Phoneme in two samples

5.3. TEXT INDEPENDENT SPEAKER VERIFICATION (TI-SV) USING GMM

With a speech signal (X) for training and a test speech signal (Y), speaker verification is a test to decide whether X was said by Y or not. In this verification process the text spoken by the speaker at training as well as testing time is not dependent on the exact words spoken i.e. it’s a text independent process. So this system is also known as text independent speaker verification [48] [50].

The Text Independent Speaker Verification (TI-SV) system takes its decision of true/false based on two theories.

T0: X is spoken by speaker Y
T1: X is not spoken by speaker Y
Likelihood Ratio (LR) is a measure to select between these two speculations. And this is given by the following equation (30).

\[
\text{score} = \frac{p(X|T0)}{p(X|T1)}
\]

- 30

if score ≥ θ; accept T0; else if score < θ; accept T1

Where \( p(X|T0) \) the probability density function for the assumption T0 is is spoken by speaker Y where as \( p(Y|H1) \) is the probability density function for the assumption T0 is not spoken by speaker Y. \( \theta \) is the threshold parameter which decides to accept or reject the speaker. Now the point of focus is to evaluate the two likelihoods \( p(X|T0) \) and \( p(Y|H1) \).

Figure 5.5 explains the way a decision of speaker verification is taken based on a threshold value.

![Figure 5.5: TI-SV system](image)

The procedure begins with de-noising the speech signal using Wiener filter followed by feature extraction. For our experimentation we have considered 13 MFCC, 13 delta- MFCC and 13 double delta- MFCC i.e. a total of vectors of 39 values for each frame (F) of a speech signal which is given the equation (31).

\[ F = \{ \vec{f_1}, \ldots, \vec{f_t} \}, \]  

- 31

Where \( \vec{f_t} \) is a feature vector of frame 't'

Then the feature vectors evaluate the likelihoods T0 and T1. Logically, a model given by \( \lambda_{hyp} \) indicates T0, which depicts the evaluated speaker Y in the feature space of \( \vec{x} \). For instance, one could recognize that a Gaussian stream best addresses the spread of highlight vectors for T0 so that \( \lambda_{hyp} \) would contain the mean vector and covariance system parameters of Gaussian dispersion.
The model $\lambda_{hyp}$ speaks to the option speculation, T1. The logarithm of probability proportion measurement is used by giving the log LR.

$$\Lambda(X) = \log p(Y|\lambda_{hyp}) - \log p(Y|\lambda_{\overline{hyp}})$$

While the model for T0 is all that much described and can be assessed using get ready talk from Y, the model for $\lambda_{hyp}$ is less all around portrayed since it possibly must identify with the entire space of possible unmistakable alternatives for the conjectured speaker. Two vital methods of insight have been taken for this decision speculation appearing. The premier theory is to utilize an orchestrated of other speaker models to cover the elective's space speculation. In unmistakable affiliations, this arrangement of various speakers has been called probability degree sets, partners, and foundation speakers. Given a course of action of N foundation speaker models $\{\lambda_1, ..., \lambda_N\}$ the alternative hypothesis model is identified with by mathematical statement.

$$p(Y|\lambda_{hyp}) = f(p(Y|\lambda_1), ..., p(Y|\lambda_N))$$

Where $f(\bullet)$ is some limit, for event, common or most persuading, of the likelihood values from the establishment speaker set. The determination, size, and mix of the establishment speakers have been the subject of much research. Taking everything in account, it has been found that to get the best execution with this procedure obliges the utilization of speaker-specific establishment speaker set.

An essential stride in the above likelihood thickness capacity thickness limit is the prodigies' choice probability limit $p(X|\lambda)$. The decision of this utmost is all around reliant on the parts being utilized and also specifics of the application. For substance free speaker insistence, where there is no past information of what the speaker will say, the best probability point of confinement has been GMMs. PDF form of speech singal is given in figure 5.6.

In speaker affirmation applications, where there is different earlier learning of the discourse substance, extra brief information can be joined by utilizing secured Markov models (HMMs) for as far as possible.
Till now, regardless, the exploitation of other jumbled likelihood limits, for occasion, those in context of HMMs, have shown no point of slant over GMMs for substance free speaker conspicuous verification tasks like in the NIST speaker affirmation evaluations (SREs).

The benefits of GMM lies in its simple usage, certainty level which can be gotten from the back probabilities. It is the quickest calculation for learning blend models. As this calculation expands just the probability, it won't predisposition the methods towards zero, or inclination the bunch sizes to have particular structures that may or may not make a difference.

5.4. GMI

To get acoustic data present in a speech signal, we constructed a speaker independent Gaussian Mixture Model. To train this model we consider TIMIT corpus as a benchmark dataset. The details of this dataset are mentioned in Chapter -3. A mixture of acoustic data is obtained as a result of unification of several speech signals in the TIMIT corpus. The process of construction of SI-GMM is shown in figure 5.7. For better classification, we use the entire dataset to train SI-GMM so that it performs well for the classification of data from which the model is just trained.
Followed by the construction of SI-GMM, based on the highest score of likelihood an index is organized in this Gaussian Mixture Indexing (GMI) method. Indexing is applied for both train and test data as shown in figure 5.8.

5.5. SPEAKER CORRELATION COMPUTATION
The main goal of the system is to fill the acoustic holes in the train as well as test data as they are sparse in nature. Precisely caution must be taken to choose minimum number of cohort speakers to fill the holes. To evaluate how close these speech signals are, we adopted likelihood based calculations.

The test speaker gets the information from the cohort speakers to fill the acoustic holes in the training speech signal. Care must be taken while choosing cohort speakers as well as minimum number of them to fill the acoustic holes. The method adopted in our experimentation to select cohort speaker is Probability-Score-Method (PS-M).

PS-M uses the likelihood to measure the closeness between the test data and the designed speaker model. It is an indirect way of measuring the similarity as it does not involve features of a speaker in its process directly. By a symmetrical evaluation, the likelihood between the speaker’s speech signals is measured, given by equation (34).

$$S_{ni} = p(X_{n}^{test} | \Lambda_{i}^{train}) + p(X_{i}^{train} | \Lambda_{n}^{test})$$

Where $$1 \leq i \leq N_{train}$$ and $$1 \leq n \leq N_{test}$$

The likelihoods mentioned in the above equations tell us the proximity between test speaker to train speaker and vice versa.

### 5.6. FUSION ILLUSTRATION OF NORMAL DISCUSSION (FIND) BASED MODELING

Languages like English have various rates of each phoneme in it. Modeling for speaker verification will be easier if the input speech signal consists of all the phonemes. The peculiar case which we have considered for our experimentation is when the speech signals if filled with acoustic holes. A solution for this can be filling of those acoustic holes using average phoneme rate data gathered from normal human discussions. This process is illustrated in figure 5.9. An unbiased acoustic token histogram is drawn from the average phoneme rate from every feature frame using GMI. This representation is called as Fusion Illustration of Normal Discussion (FIND) table.
FIND Algorithm is as follows:

**Step 1:** calculate the rate of occurrence of every mixture indexing information for each training speech signal’s information $i$, $1 \leq i \leq N_{dev}$, by Gaussian Mixture Indexing. Find table is build using this information.

**Step 2:** Gather equal number of same mixture token feature frames using FIND table. Based on the amount of data required for filling and the amount cohort data present for each feature, the numbers of top cohort speakers for every testing speaker vary.

**Step 3:** For every test speech signal, build a FIND cohort GMM based on gathered information from collected.

**Step 4:** Using MAP adaptation get the ultimate speaker model for the test speech signal from the basic cohort model.

To fill the acoustic holes with cohort speaker’s phonemes we need sufficient as well as suitable data. From the data present in the FIND table, it evaluates the amount of each phoneme set required for filling the acoustic holes. The last step of our model is to get effective cohort speakers to fill the holes.
5.7. RESULTS

For the experimentation, datasets mentioned in chapter -3 are utilized. Attributes for execution of is as specified in chapter -4. The results of the system are as shown in tables 8 and table 9. They clearly depict the impact of sparse data in the execution process. When compared to the results of the conventional models for speaker verification for sparse data, the proposed system with derived feature selection algorithm placed in it have shown better results.

Table 5.1: DCF values when experimented with two different datasets

<table>
<thead>
<tr>
<th>Number of Gaussians</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone Conversation 5db noise</td>
<td>0.0236</td>
<td>0.0274</td>
</tr>
<tr>
<td>10db noise</td>
<td>0.347</td>
<td>0.434</td>
</tr>
<tr>
<td>SR cum ER dataset</td>
<td>0.0145</td>
<td>0.0186</td>
</tr>
</tbody>
</table>

Table 5.2: EER values when experimented with two different datasets

<table>
<thead>
<tr>
<th>Number of Gaussians</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone Conversation 5db noise</td>
<td>2.61</td>
<td>2.92</td>
</tr>
<tr>
<td>10db noise</td>
<td>3.32</td>
<td>3.67</td>
</tr>
<tr>
<td>SR cum ER dataset</td>
<td>1.89</td>
<td>2.83</td>
</tr>
</tbody>
</table>
Results of text independent speaker verification system with MOHABC algorithm placed as part in it are shown in figures 5.10 & 5.11. Here we have considered sparse data from Speech Recognition cum Emotion Recognition dataset and Telephone Conversation with a noise introduced in it.
5.8 ASSESSMENT

Further comparison was conducted with the following existing systems.

Angkititrakul, P. (2007) used a simple decision rule for text independent speaker recognition with less duration of test and train speech data. When TIMIT was used 9%-23% of an average EER was obtained.

Jun-Won Suh et.al. (2012) in his paper proposed a new methodology of ranking of acoustic phonemes to fill the sparse speech signal. Such a system was compared with human listener. The machine performed with an accuracy of 95% (System-1).

Nicolas Scheffer et.al. (2014) proposed a system using deep belief networks on phonetic information from short duration speech signal for speaker verification. Experimentation showed an accuracy rate on an average of 50% (System-2).

Yi Nie et.al. (2014) focus was on the influence of sparse data on speaker verification. K-SVD algorithm, generalization of K-means algorithm was used to train dictionary with Singular Value Decomposition (SVD). The proposed sparse representation showed an average Error Equal Rate (EER) of 14.23%

Chart representation of the above results is shown in figure 5.12.

Figure 5.12: Result Analysis
5.8. SUMMARY

In this chapter we presented the third contribution towards thesis. When sparse data of 2~5 sec is provided for speaker verification, we implemented a combination of MOHABC algorithm for optimization of features and GMM Mixture Indexing to improve the verification accuracy. GMM Mixture Indexing is a technique where feature frame of the speaker training data is labeled with the highest probability in a table known as Gaussian Mixture Index (GMI). The speaker feature frame is classified into one Gaussian mixture index in the GMI scheme and this indexing process is applied to both the trained data set and test data. The indexed information from the train data is used for the speaker adaption procedure, which used different adaption data depending on the ranking of the indexed information. To fill acoustic holes for sparse training data, the test speaker borrows data from acoustically close speakers. When acoustically similar data are selected from cohort speaker data, caution should be exercised to ensure a minimum number of cohorts for filling acoustic holes. The method adopted to select similar characteristic speakers is Probability-Score-Method (PS-M). It uses the probability of the feature frame against the speaker model. The average phoneme occurrence is counted as a unit in the indexed mixture of each speech feature frame using the GMI. A balanced acoustic token histogram is formed from this information, and is called the FIND table.

We considered speech signals of very less duration like 2 ~ 6 sec for training as well as testing. For mapping we used MAP adaptation and the benchmark dataset considered is TIMIT corpus. Using the phonemes from the cohort speech signal we can fill the holes present in the sparse speech signal. To achieve this task we have come up with a new procedure where we index the phonemes of speech signals in an indexing table. When the proposed feature selection algorithm was implanted as a part of this verification process, the results surpassed the efficiency of conventional methods with an average accuracy of 95.25%