1. Introduction

In the digital era authentication and security have become an important issue. With the exponentially increasing use of services like phone banking, internet banking, online shopping or voice mail, the traditional methods of authentication like passwords are not sufficient. Hence Biometric authentication techniques are being explored. In the biometric technique some behavioral trait or part of human body is used for authentication. There are several categories of biometrics: fingerprints, palm prints, hand geometry, retina, iris, face, handwriting, voice etc. All these categories have certain advantages and disadvantages over each other. Voice as a biometric has distinct advantages over other techniques like [1]:

- it does not require special hardware, so implementation cost is low
- it is easy to use and easily accepted by users
- users can authenticate remotely
- enrollment is faster and easier
- authentication is faster

Along with these advantages it does have certain drawbacks like:

- security
- changes in human voice over time

Speaker Recognition systems use voice as a unique characteristic that can identify a person [2 - 4]. We as human beings can recognize a person whose voice we have heard before, with a fairly good accuracy. The aim of speaker recognition systems is to imitate this capability of humans. The voice or speech signal can be used as a unique trait to identify a person.
1.1 Classification of Speaker Recognition Systems

The classification of the Speaker Recognition Systems \([3 – 6]\) is shown in Fig.1.1. Speaker Recognition Systems can be divided into Speaker Verification and Speaker Identification. Speaker Identification determines which registered speaker provides a given utterance from amongst a set of known speakers, also known as 1:N matching. Speaker verification accepts or rejects the identity claim of a speaker, also known as 1:1 matching. Speaker Identification task can also be classified into closed set and open set Speaker Identification \([5, 7]\). In closed set problem, from N known speakers, the speaker whose reference template has the maximum degree of similarity with the template of input speech sample of unknown speaker is obtained. This unknown speaker is assumed to be one of the given set of speakers. Thus in closed set problem, system makes a forced decision by choosing the best matching speaker from the speaker database. In the open set text dependent speaker identification, matching reference template for an unknown speaker’s speech sample may not exist.
Speaker Identification task can be further classified into text-dependent or text-independent task [5, 6]. In the former case, the utterance presented to the system is known beforehand. In the latter case, no assumption about the text being spoken is made, but the system must model the general underlying properties of the speaker’s vocal spectrum. In general, text-dependent systems are more reliable and accurate, since both the content and voice can be compared [4, 5].

1.2 The general Speaker Identification Process

The General scheme for Speaker Identification [74, 75] is shown in Fig. 1.2. Test and reference patterns (feature vectors) are extracted from speech utterances statically or dynamically. At the training stage, reference models are generated (or trained) from the reference patterns by various methods. A reference model (or

![Fig. 1.2: Speaker Identification Process](image-url)
reference templates at the pattern matching stage. The comparison may be conducted by similarity measure using either distance or statistical parameters. After comparison, the test pattern is labeled to a speaker model at the decision stage. The labeling decision is generally based on the minimum risk criterion. The Speaker Verification process is shown in Fig.1.3. In the verification process the feature vectors extracted from the test speaker (particular ID) is matched with his own template stored in the database. Matching is done based on the similarity measure. The decision to accept or reject a speaker depends on the threshold.

![Fig. 1.3: Speaker Verification Process](image)

### 1.2.1 Feature Extraction

Feature extraction is the process of extracting features, which are representative values that can uniquely characterize a person’s voice, from the speech signal. These representative values should have the following properties [8 - 11]:

1. High inter-speaker variation,
2. Low intra-speaker variation,
3. Easy to measure,
4. Robust against disguise and mimicry,
5. Robust against distortion and noise,
6. Maximally independent of the other features.

The first two requirements state that features should be discriminative as possible. The next property states that features should be easily measurable. This includes two factors. Firstly, the feature should occur frequently and naturally in speech so that it could be extracted from short speech samples. Secondly, the feature extraction itself should be easy. A good feature is robust against several factors like voice disguise, distortion, noise and data acquisition system. Finally, different features extracted from the speech signal should be maximally independent of each other. If two correlated features are combined, nothing is gained, and in fact, this may even degrade recognition results. The ideal recognition system which satisfies all of these properties is difficult to realize.

Feature extraction methods of speaker recognition given in literature use LPC (Linear Predictive Coding) [12], MFCC (Mel Frequency Cepstral Coefficients) [13 – 16] and wavelets [17 – 19].

1.2.2 Feature Matching

In this phase, the feature vector extracted from the test sample is compared and matched with the reference features stored in the database. There are two broad approaches [11]: parametric (stochastic) and non-parametric (template) approaches for speaker matching. In the parametric approach, a certain type of distribution is fitted to the training data by searching the parameters of the distribution that maximize some criterion. The non-parametric approach, on the other hand, makes minimal assumptions about the distribution of the features. Feature matching has been done using Vector Quantization, a non-parametric approach [20 – 24], HMM
Speaker Identification using Orthogonal Transforms and Vector Quantization

(Hidden Markov Model) [24 – 25], GMM (Gaussian Mixture Model) [26], which are parametric approaches. The pattern matching phase consists of computing a similarity score for the unknown speaker’s feature vectors and all speaker models. The similarity (or dissimilarity) measure depends on the type of the speaker models. The most commonly used distance measure is the Euclidean Distance (ED) or the squared Euclidean distance [2]. We have used one more distance measure Manhattan Distance (MD) also and comparative performance of both is given.

1.2.2.1 Similarity measures

For this work two distance measures have been used and the performance of both has been compared. The Minkowski distance [65] is a metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. The Minkowski distance of order p between two points

\[ P = (x_1, x_2, \ldots, x_n) \text{ and } Q = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \]

is defined as \( d_{pq} \)

\[ d_{pq} = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \]  \hspace{1cm} (1.1)

Minkowski distance is typically used with \( p \) being 1 or 2. The latter is the Euclidean distance, while the former is known as the Manhattan distance.

1. Manhattan Distance

Manhattan distance (MD) [28] is defined as the Minkowski distance of the order 1 or 1-norm distance (where \( p=1 \)). The 1-
norm distance is called the taxicab norm or Manhattan distance, because it is the distance a car would drive in a city laid out in square blocks (if there are no one-way streets). In n dimensions, the MD between two points P and Q is given by eq. (1.2), where $x_i$ (or $y_i$) is the coordinate of P (or Q) in dimension i.

$$d_{PQ} = \sum_{i=1}^{n} |x_i - y_i|$$

(1.2)

2. Euclidean Distance

Euclidean distance is defined as the Minkowiski distance of the order 2 or 2-norm distance (where $p=2$). Euclidean Distance (ED) [27, 28] is defined as the straight line distance between two points. It is what would be obtained if the distance between two points were measured with a ruler: the "intuitive" idea of distance. In n dimensions, the ED between two points P and Q is given by eq. (1.3), where $x_i$ (or $y_i$) is the coordinate of P (or Q) in dimension i.

$$d_{PQ} = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

(1.3)

The results obtained by using these two distance measures for accuracy as well as performance measurement of different algorithms has been given in subsequent chapters.

1.2.3 Decision Making

The final step in speaker recognition process is the decision. The feature extraction and pattern matching are same for different
speaker recognition tasks, but the decision depends on the task: Identification or Verification. Let us denote generally a speaker model of speaker $i$ by $S_i$, and let $S = \{S_1, \ldots, S_N\}$ be the speaker database of $N$ known speakers. Without assuming a specific speaker model/classifier, let $\text{score}(X, S_i)$ be the match score between the unknown speaker’s feature vectors $X = \{x_1, \ldots, x_T\}$ and the speaker model $S_i$. In the case of distance based classifiers, minimum distance corresponds to best match. In closed-set speaker identification task, the decision is simply the speaker index $i$ that yields the minimum distance, where $i$ is given by eq. (1.4).

$$i = \min_i \text{dist}(X, S)$$

where the minimum is taken over the speaker database $S$. In the open set identification task, the decision is given as given by eq. (1.5).

$$\text{dist}(X, S_i) \begin{cases} < \Theta_i, & \text{accept} \\ \geq \Theta_i, & \text{reject} \end{cases}$$

Where $\Theta_i$ is the threshold. The threshold can be set the same for all speakers, or it can be speaker-dependent. The threshold is determined so that a desired balance between the two types of errors False Acceptance Rate (FAR) and False Rejection Rate (FRR) [5, 64].

1. **False Rejection Rate (FRR)**

The rate, generally stated as a percentage, at which authentic, enrolled persons are rejected as unidentified or unverified persons by an identification system is termed the false reject rate. False rejection is sometimes called a Type I error [5]. In access control, if
the requirement is to keep the “bad guys” out, false rejection is considered the least important error. However, in other biometric applications, it may be the most important error. When used by a bank or retail store to authenticate customer identity and account balance, false rejection means that the transaction or sale (and associated profit) is lost, and the customer becomes upset. Most bankers and retailers are willing to allow a few false accepts as long as there are no false rejects.

2. False Acceptance Rate (FAR)

The False Acceptance Rate, generally stated as a percentage, is the rate at which un-enrolled or impostor persons are accepted as authentic enrolled persons by an identification system. False acceptance is sometimes called a Type II error [5]. This is usually considered to be the most important error for a biometric access control system. FRR and FAR are given by eq. (1.6) and eq. (1.7) respectively.

\[
\text{FRR} = \frac{\text{true claims rejected}}{\text{total true claims}} \times 100 \quad (1.6)
\]

\[
\text{FAR} = \frac{\text{imposter claims accepted}}{\text{total imposter claims}} \times 100 
\]

\[\text{GAR} = 100 - \text{FRR} \quad (1.8)\]

\text{GAR} given by eq. (1.8) is defined as the \textbf{Genuine Acceptance Rate (GAR)}, in percentage.

Thus FAR is the error with which an imposter is accepted and FRR is the error with which a genuine or true speaker is rejected. There is a trade-off between the two errors. When the decision threshold \(\Theta_i\) is increased: FAR increases but FRR decreases, and vice versa. The balance between these two depends on the application. Since
either of the two types of errors can be reduced at the expense of an increase in the other, a measure of overall system performance must specify the levels of both types of errors. The tradeoff between FAR and FRR is a function of the decision threshold. FAR and FRR are plotted against the decision threshold. The point of intersection of these two curves is defined as the Equal Error Rate (EER). The EER is the value for which the FAR and FRR are equal. The system performance can be given by Performance index (PI), which is defined as given by eq. (1.9).

\[
\text{PI} \, (\%) = 100 - \text{EER} \, (\%)
\]  

(1.9)

1.3 Different Databases (Corpora) for Speaker Identification

A list of some of the publicly available corpora for speaker recognition is given in Appendix I. The primary suppliers of these corpora are the European Language Resources Association (ELRA) [29], the Linguistic Data Consortium (LDC) [30], and the Oregon Graduate Institute (OGI) [31], as indicated next to each corpora name. Out of these databases, evaluation has been done on the last two databases:

1. **CSLU**, which is a text dependent Speaker Recognition database obtained from OGI. The CSLU Speaker Recognition Database consists of telephonically recorded speech spanning twelve collected over a two year period. Table 1.1 shows the database description. These speech signals have very low amplitude range. These signals are scaled up to the level ‘-1’ to ‘+1’. Also preprocessing was done to remove the long silent parts in between the words.
2. **Local Database**, which has been prepared locally. The speech samples used in this work are recorded using Sound Forge 4.5. The sampling frequency is 8000 Hz (8 bit, mono PCM samples). Table 1.2 shows the database description. The samples are collected from different speakers. Five iterations of four different sentences of varying lengths are recorded from each of the speakers. Twenty samples per speaker are taken. For text dependent identification, four iterations of a particular sentence are kept in the database and the remaining one iteration is used for testing. These speech signals have an amplitude range of ‘-1’ to ‘+1’.

The detailed description of these two databases is given in Appendix I.

<table>
<thead>
<tr>
<th>Table 1.1 Description of CSLU Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>No. of Speakers analyzed</td>
</tr>
<tr>
<td>Speech type</td>
</tr>
<tr>
<td>Recording conditions</td>
</tr>
<tr>
<td>Sampling frequency</td>
</tr>
<tr>
<td>Resolution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 1.2 Description of Local Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>No. of Speakers</td>
</tr>
<tr>
<td>Speech type</td>
</tr>
<tr>
<td>Recording conditions</td>
</tr>
<tr>
<td>Sampling frequency</td>
</tr>
<tr>
<td>Resolution</td>
</tr>
</tbody>
</table>

### 1.4 Problem Formulation

In this thesis, different approaches for Text Dependent Speaker Identification in closed set and open set environment are studied
and various algorithms are introduced. For the purpose of evaluation, two databases are used: Local Database and CSLU Database.

Work on Speaker Identification in this thesis can be broadly divided into two domains:

- **Transform Domain**
  
  In the Transform Domain, the speech signal is processed in two ways:
  
  - Full Transform of the 1 Dimensional speech signal
    - The techniques proposed here are Sectorization [61 - 63] and Amplitude Distribution [66 - 68]
  
  - Column Transform of the frames of speech signal
    - The techniques proposed here are Row mean of Spectrogram [71] and Row Mean of Column Transform [69, 70]

- **Vector Quantization**
  
  Here Vector Quantization has been studied in spatial and Transform Domain
  
  - In the spatial domain, VQ has been analyzed [72 - 74] in 16 and 32 dimensional vector space with 0%, 25% and 50% overlap between the training vectors. Five different clustering algorithms LBG, KPE, KEVR, KMCG and KFCG are used for evaluation and comparison of results.
  
  - In the Transform Domain, four different Transform techniques [75]: DFT, HT, DCT and DST are applied on the framed speech signal and the magnitude of the transformed framed forms the training vector. Here again the five clustering algorithms mentioned above are evaluated.
For the Transform domain, the results are evaluated for the traditional approach of MFCC - LBG technique and the same is extended to the other four clustering algorithms.

1.5 Scope/Applications

Speaker Identification systems have a wide number of applications like:

- Controlled access to restricted areas, facilities and equipment
- Access control to databases (e.g. telephone and banking application)
- Secured use of access cards (e.g. calling and credit cards)
- Electronic commerce (Online shopping)
- Voice mail
- Information and reservation services
- Remote access to computer networks
- As an additional parameter for speech recognition systems

The proposed work can be very well used in all of the above mentioned applications.