CHAPTER 2

LITERATURE REVIEW OF SPEAKER RECOGNITION, VERIFICATION AND AUTHENTICATION

Chapter overview:

- MODEL OF SPEECH RECOGNITION APPROACHES TO SPEECH RECOGNITION
  - Acoustic Phonetic Approach (APA), Hidden Markov Model (HMM)
  - Template Based Approach (TBA), Dynamic Time Warping (DTW)
  - Stochastic Approach (SA), Vector Quantization (VQ)
  - Pattern Recognition Approach (PRA)
  - History of Speech Recognition, Speaker Identification and Authentication:
- WHY IS THE BODO SPEAKER AUTHENTICATION SYSTEM NECESSARY?

CHAPTER 2

REVIEW OF SPEAKER RECOGNITION, VERIFICATION AND AUTHENTICATION

2.1 Speech Recognition

Speech is the most predominantly accepted, efficient and natural way for human beings to communicate. It is, therefore, sensible to investigate, develop and deploy technologies that facilitate speech-enabled human computer interaction, in environments where users may experience efficiency and convenience (Gaikwad, Gawali & Yannawar, 2010:24-28) [119]. Speech-enabled human computer interaction in access control can be an area of focus, where users will find traditional identification and authentication methods less convenient or physically intrusive.

When a machine enabled system or program acquires the capability to recognize words/phrases/sentences of any language used for communication and can transforms the same to a machine understandable form is called Speech Recognition. For last few decades researchers have been showing a great amount of enthusiasm to develop Automatic SR. With the IT tools are reaching the hooks and corners of everyday life, attraction towards targeting to communicate to machines through spoken language is obviously gaining momentum among the research community. The idea of using Natural language as an interface between human and machine can be achieved with SRS. Prime target of SRS is to design and devise set of techniques and systems to feed machines with spoken voice. Remarkable achievements have been reached by the research community in last few decades. Statistical modeling of speech has reached the pinnacle which have resulted in ASRS (automatic speech recognition system) finding grounds in application for

- human machine interface
- automatically processing phone calls in speech networks
- information systems which works through, by answering queries and updating clients/users with several spheres of knowledge such as access to database (like banking, travel, aviation, automobile portal, speech transcription etc), travel/stock/weather/data entry/voice detection, differently able community service, airlines/railway reservation, voice command systems as a replacement for traditional menu based programs etc.
- Even highly configured fighter jets, traffic controllers training etc.

Speech recognition focuses on what is being said. A speech recognition system typically translates the spoken message into machine understandable language. This, therefore, enables computers to understand and respond to voice commands (Juang and Chen, 1998:24-48) [17]. Siri is different to Windows speech recognition in that it recognises natural speech without the need to provide special commands, such as in the case of many other traditional speech recognition systems (Apple.Inc, 2012). A final example of a speech recognition implementation is the flight information system deployed by the Airports Company South Africa in 2004. This system enabled people to find out certain flight information through a spoken telephonic conversation with an automated information provider (FHC, 2004).

Speaker recognition is a biometric modality concerned with determining who the speaker is. It is not as common as speech recognition and differs by not attempting to draw any meaning from the message contained in what is said (NSTC Subcommittee of Biometrics, 2006). Speaker recognition extracts and compares the unique physiological and behavioural features from the speaker's voice in order to

determine or verify the speaker's identity (Reynolds, 2002:4072-4075) [90]. This capability has motivated the incorporation of speaker recognition in security applications, such as access control (Doddington, 1985:1651-1664). One fiction example to help understand practical speaker recognition enabled access control system may be mentioned here. In the movie "Sneakers" (1992), Werner Brandes, played by Stephen Tobolowsky, uses the phrase "Hi, my name is Werner Brandes. My voice is my passport. Verify Me.", in order to gain physical access to his office using nothing but his voice (Matthieu, 2011). In speech recognition process, aim is to get the optimal word sequence within linguistic constraints. Spoken utterances consist of key elements namely phonemes/syllables/words. ASRS understands a sentence as a combination of such units. The acoustic models gives SRS acoustic clues coming from these fundamental units which in tandem with proper sentence forming rules to produce reach sentences to acquire the hypothesis sentence. That means, in SRS, pattern matching mechanism involves two aspects: *acoustic/phonetic* and *symbolic/graphic*. As far as acoustic domain is concerned, for all pre existing classes the feature vector of the test sample speech wave (generally small section) is matched. Now this small test speech segment gets the identity of the class having maximum likely score. By performing this procedure all the feature vectors are assigned their class labels. After acquiring these labels (a lattice of label), hypothesis is processed along with the language model and the recognized sentence is produced.

Table 2.1: Linguistic levels and their corresponding object of study.

LEVEL	STUDY OBJECT
Phonetics	Human speech wave physical properties
Phonology	Language specific (or across the language) sound system
Morphology	Word's internal structure
Lexicon	Language specific words and phrases
Syntax	The structure of grammatical sentences governing rules
Semantics	Meaning of words and phrases
Pragmatics	Utterances used for communication
Discourse analysis	Text based analysis of language (spoken and written or signed)

Based on utterance pattern, SRS are categorizes as follows:

- ✓ *Isolated phoneme or word, Connected words, Continuous* utterance

A single utterance or word at an instance is feed to isolated SRS meaning that gaps are maintained between spoken words. It will carry "Listen/not-Listen" phases and utterer pauses (generally 100-250 msec) after each utterance. In case of continuous SRS users speaks in a natural way (generally conversation of 150-250 words/minute without following any speaking style) and content understanding part is performed by the computer. This is more complex job in comparison to isolated word SRS as word boundary is undefined. Finally the connected word system resembles isolated word system. Though 'run together' of different speech are allowed (with small pauses in the middle). This is a balanced combination of previous two, with the pre condition that the utterer requires no pause in between at the same time must take care of clear pronunciation and adequate stress in individual words.

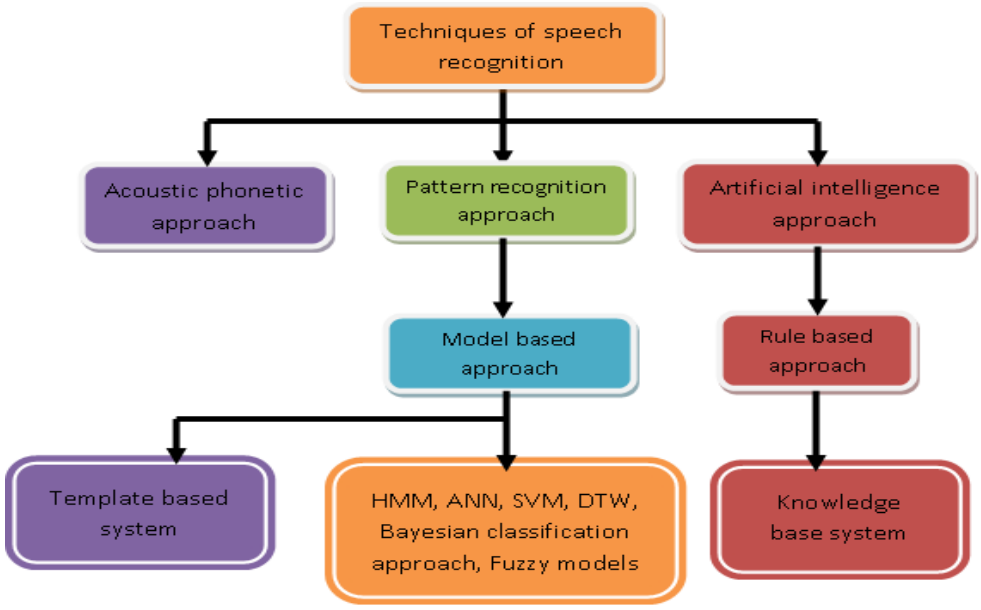

Figure 2.1: Speech recognition classification.

2.2 Model of Speech Recognition

The speaker recognition problem is one that has baffled scientists for over half a century and still continues to do so. It remains an active area of research (Furui, 2005:1-9) [117]. Considering a trained SR model with a test sample speech wave, our target is to hypothesis the most likely sentence (in a form of sequence of words). Assuming A as the acoustic feature series computed from the sample speech, the SRS should return the best matching value $\hat{w}$ such that:

$$\hat{w} = \arg\max_{w} p(w \mid A) \qquad \text{--- 2.1}$$

After applying Bayes' rule, $p(w \mid A)$ will become:

$$p(w \mid A) = \frac{p(A \mid w)P(w)}{P(A)} \qquad \text{--- 2.2}$$

Here, $p(w \mid A)$ represents likelihood ratio of A for w. $p(w)$ stands for probability calculated from the language model. $p(A)$ represents an independent

priori probability of A (neglected in the maximization operation). So, word sequence probability is the multiplication of the probabilities $p(A|W)$ [for acoustic model] and $p(W)$ [language model]. We can represent this through Figure 2.3 given below.

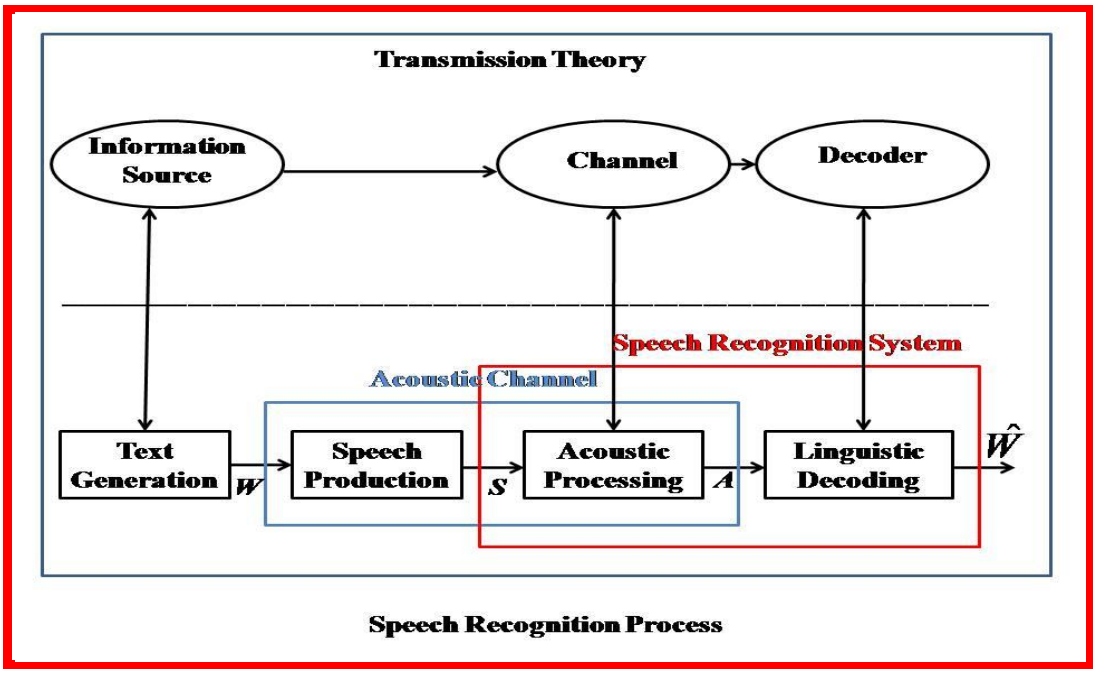

Figure 2.2: Information transmission theory for SAS with Statistical modeling framework.

2.3 Approaches to Speech Recognition:

There are three basic approaches of speech recognition. They are-

2.3.1 Acoustic Phonetic Approach

2.3.2 Pattern Recognition Approach

2.3.3 Artificial Intelligence Approach

2.3.1 Acoustic Phonetic Approach:

The earliest approaches to speech recognition were founded on evaluating speech sounds and arranging proper labeling of all sounds. It establishes foundation of the acoustic phonetic approach (Hemdal and Hughes 1967). The proposition is that there exist finite, distinctive phonemes (phonetic units) in spoken language and they are largely characterized by a set of acoustics properties that are manifested in the speech signal in time domain. The acoustic properties varies widely-

- ✓ Due to speaker to speaker differences and
- ✓ Due to environmental effects (normally termed as co articulation effect).

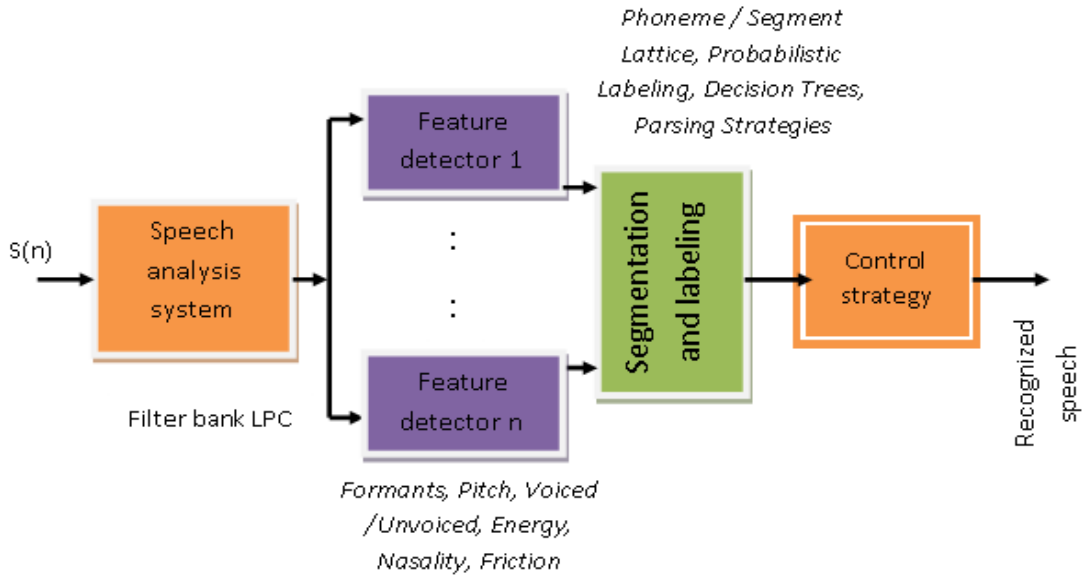

Figure 2.3: Block diagram of acoustic phonetic speech recognition system.

However it is presumed that the variability governing terms are straightforward and machine adapts with ease. The Figure 2.3 depicts the block diagram representing the same. The first step in the acoustic phonetic approach is a *spectral analysis* of the speech that depicts the broad acoustic properties of the different time varying speech units. The most common techniques of spectral analysis are the class of filter bank methods and the class of linear predictive coding (LPC) methods. The next step is

the *feature- detection* stage where the spectral measurements are converted to a set of features that describe the broad acoustic properties of the different phonetic units. The expected features for recognition are *nasality* (presence or absence of nasal resonance), *frication* (presence or absence of random excitation in speech), *formant locations* (frequencies of the first three resonances), *voiced/unvoiced classification* (periodic or aperiodic property), and *ratios of high and low frequency energy*. The next step is a *segmentation and labeling phase* in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech utterance. This stage is the core of the acoustic phonetic recognizer which is the most difficult one to carry out reliably. Different control strategies are used to limit the range of segmentation points and label opportunities. The outcome of the segmentation and labeling step is a phoneme lattice from which a lexical access procedure determines the best matching word or sequence of words [107, 109]. There are many problems that are associated with the acoustic-phonetic approach. These problems are as listed below:

- This method requires extensive knowledge of the acoustic properties of phonetic units.
- The choice of features is considered mostly based on ad hoc ways. For most systems the choice of features is based on perception and is not optimal in a well-defined and meaningful sense.
- Ad-hoc method generally uses binary decision trees such as CART method to make the decision trees more robust. But since the choice of features is most likely to be sub-optimal, optimal implementation of CART is rarely achieved.

- No well-defined, automatic procedure exists for tuning in the labeling method.

There is not even an ideal way of labeling the training speech in a consistent manner and agreed on uniformly by a wide class of linguistic experts.

2.3.2 Pattern Recognition Approach

The pattern-recognition approach to speech recognition is basically one in which the speech patterns are used directly ignoring feature determination (in the acoustic phonetic sense) followed by segmentation. It involves two fundamental steps namely, *training of speech patterns* and *recognition of patterns* (comparing patterns) (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993). Following block diagram displays a clear view.

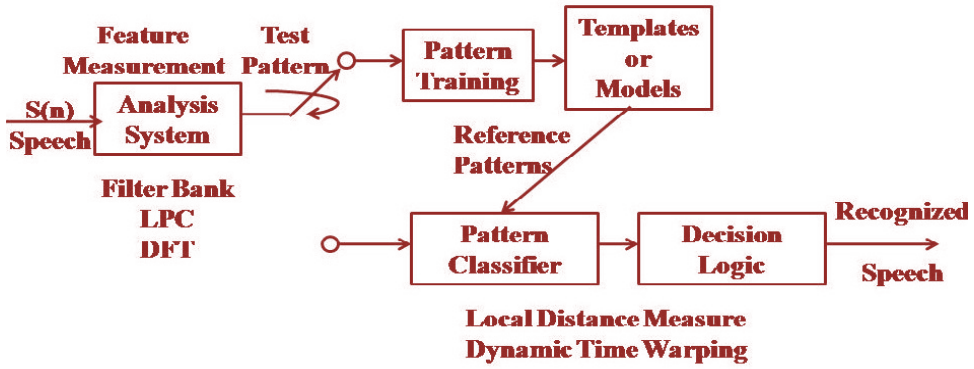

Figure 2.4: Block diagram of pattern-recognition speech recognition system.

Prime feature of this approach lies in application of smoothly formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. This type of characterization of speech via training is called pattern classification because the machine learns which acoustic properties of the speech class are dependable and repeatable across all training tokens of the pattern

[30, 33]. Speech patterns are represented through speech templates or a statistical model (like HMM) and can be applied to a sound (smaller than a word, a word, or a phrase). The pattern comparison plays the most crucial act in the success of the system because it performs a direct comparison of the unknown speech. In the comparison stage, a direct matching is performed between the speech to be recognized (unknown) with each possible pattern learned in the training stage aiming to find the identity of the unknown speech based on the match score of the patterns. The approach has gained dominance in speech recognition. This method is used widely for the following reasons. They are:

- Easy to use and easy to understand. It poses mathematical and communication theory justification as individual procedures used in training and decoding.
- Robust and invariant to different aspects (e.g. algorithm applied, decision rule used, vocabulary set, different users, selected feature sets, decided pattern etc). Thus it can be applied to a wide range of speech units/words/vocabularies, background environments, transmission medium/conditions etc.
- It provide high performance on any task that is reasonable for the technology, and provides a clear way to extend the technology in a wide range of directions such that the performance degrades elegantly as the problem grows more and more complex.

Some of the well known pattern recognition approaches/models are discussed here:

2.3.2.1 Template Based Approach:

Template matching is one of the simplest and earlier approaches. A family of techniques that have advanced the field of SR considerably during the last few decades is Template based approach to speech recognition. The logic is classical. Prototype speech patterns are collected (recorded) as reference patterns that represent the dictionary of registered speaker's words. A matching of an unknown spoken utterance to the reference templates is performed and best matching pattern is found. Templates for entire words are constructed as far as possible. It provides resistance to faults resulted from segmentation/classification of smaller but acoustically more changeable units like phonemes. Accordingly, each word must have its own full reference template. It makes template preparation and matching prohibitively expensive and impractical with the increase in the number of words beyond a few hundreds. A fundamental logic used here is that first manipulate a set of speech frames for any pattern/word using averaging procedures based on the use of local spectral distance measures to compare patterns. The other trick is to use dynamic programming to temporarily align patterns to keep track of the differences in speaking rates across speakers and for repetition of the word by the same utterer.

2.3.2.2 Stochastic Approach:

In the statistical approach, each pattern is denoted in terms of d-features or measurements and is viewed as a point in a d-dimensional space [59]. The aim is to choose those features which allow pattern vectors belonging to different categories so that it can occupy compact and mutually disjoint areas in a d-dimensional feature space. Probabilistic models are brought into act for dealing with uncertain or incomplete information by this approach. Confusable sounds, speaker changing,

contextual effects, and homophone etc may cause uncertainty and incompleteness. Stochastic models have emerged suitable approach to speech recognition to sole these factors. Hidden Markov modeling (HMM) is the most popularly used one. HMM is more general and mathematical foundation is stronger compared to the template based approach. Template based model cab be described as the HMM having continuous density and identity covariance matrices possessing a slope constrained topology. Though templates training requirement is lower; still they will need the probabilistic formulation of full HMMs (and typically underperform HMMs). HMM allows assimilation of knowledge sources into a compiled architecture easily compared. HMM's flip side is that it does not give primary focus on the recognition process. So for performance enhancement, errors investigation of an HMM system becomes tough.

2.3.2.2.1 Hidden Markov Model:

Hidden Markov models (HMMs) are a ubiquitous tool for modeling time series data. The HMM for speech recognition got the popularity after the publication by S.E. Levinson, L.R. Rabiner, M.M. Sondhi and J.D. The famous HMM model is based on the years old mathematical model known as *Markov Chain*. A Hidden Markov Model (HMM) is tool which is used to determine the probability distributions over sequences of observations [94, 161]. HHM is one of the most popular approaches in speech and speaker recognition system. HMM is applied in speech, handwriting, signature, gesture recognition and medical field such as bioinformatics and genomics. This was initially introduced by Institute for Defense Analysis (IDA) in Princeton. Later on adoption of HMM by a large number of researchers, the constraint on the form of the density functions entailed a limitation on the performance of the system for instance speaker independent consideration. The

speech parameter distribution was not perfectly modeled by a simple log-concave/elliptically symmetric density function in early stage. At the Bell Laboratories (1980's), it was established significantly by extending the theory of HMM to mixture densities for the purpose of ensuring the accuracy of recognition (for speaker independent in particular and huge vocabulary speech recognition).

The name HMM got derived from two properties.

1. The HMM assumes that the observation at time *t* was generated by some process whose state S_t is hidden from the observer.
2. The HMM assumes that the state of this hidden process satisfies the Markov Property. It states that for given the value of S_{t-1}, the current state S_t is independent of all the states prior to *t-1*. So we may conclude that the state at some time encapsulates all we need to know about the past of the process in order to predict the future of the process. similarly outputs too satisfy a Markov property with respect to the states.

A state diagram of HMM is depicted in following Figure 2.5.

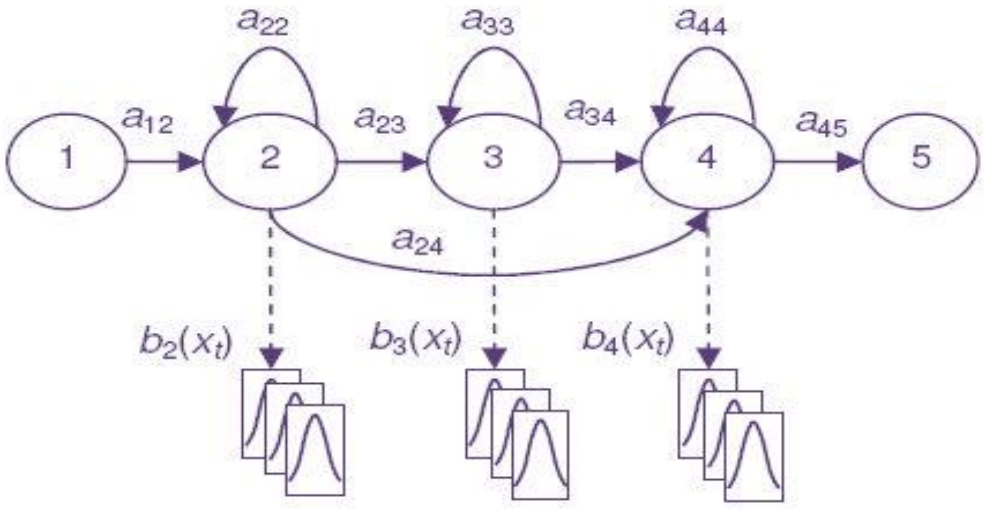

Figure 2.5: Hidden Markov model state diagram.

2.3.2.3 Dynamic Time Warping (DTW):

DTW represents an algorithm to measure similarity between two sequences though they may differ in time or speed. Dynamic time warping is a time series alignment algorithm which was originally developed for speech recognition [112]. DTW targets to align two sequences of feature vectors by warping the time axis iteratively until an optimal match between them is achieved. It has found application to video, audio, and graphics. We can in general say that, any data may be analyzed applying DTW that can be represented in linear representation. The available sequences are "warped" in a non-linear manner against the time dimension to find a similarity measure independent of certain non-linear variations in the time dimension. This type of sequence alignment method is applied in the context of HMM too. The monotonicity of the mapping in the time dimension is a restrictions found on the matching of the sequences in DTW. Continuity attracts comparatively lower attention in DTW than in other pattern matching algorithms. This algorithm is inherently suited to matching sequences with missing information, if long enough segments for matching to occur are available. As the optimization process is done through dynamic programming, it gets the name DTW. Continuity attracts comparatively less importance in DTW than in HMM, PM, VQ, ANN etc.

2.3.2.4 Vector Quantization (VQ):

Another frequently applied approach in ASR system is Vector Quantization. It is particularly useful for speech coders, meaning efficient data reduction. The transmission rate is not a major issue for ASR, so values lies in the efficiency of using compact codebooks for reference models and codebook searcher that replaces costlier evaluation methods. Each of the vocabulary word has its associated VQ codebook, as per the training sequence of several repetitions of a certain word.

Codebooks calculate the test speech and the word whose codebook gives the lowest distance measure is decided by the ASR. In basic VQ, codebooks have no explicit time information (in each word the temporal order of phonetic segments and their relative durations are ignored), since codebook entries are not ordered and can come from any part of the training words [63, 186]. Moreover, codebook entries are selected targeting average distance decrement among entire training frames (and a frame corresponding to longer acoustic segments like vowels occurs more in the training data) some indirect durational indications are accounted. Hence these segments are more likely to specify code words than less frequently occurring consonant frames, with small codebooks in particular. Nevertheless there exist code words corresponding to the constant frames because they may otherwise result in large frame distances to the codebook. Normally limited number of code words suffices to represent large volume of frames during relatively steady sections of vowels, which allows representing short, dynamic portions of the words by more codeword. It means VQ puts relative emphasis on speech transients, and advantageous for vocabularies of similar words over other methods.

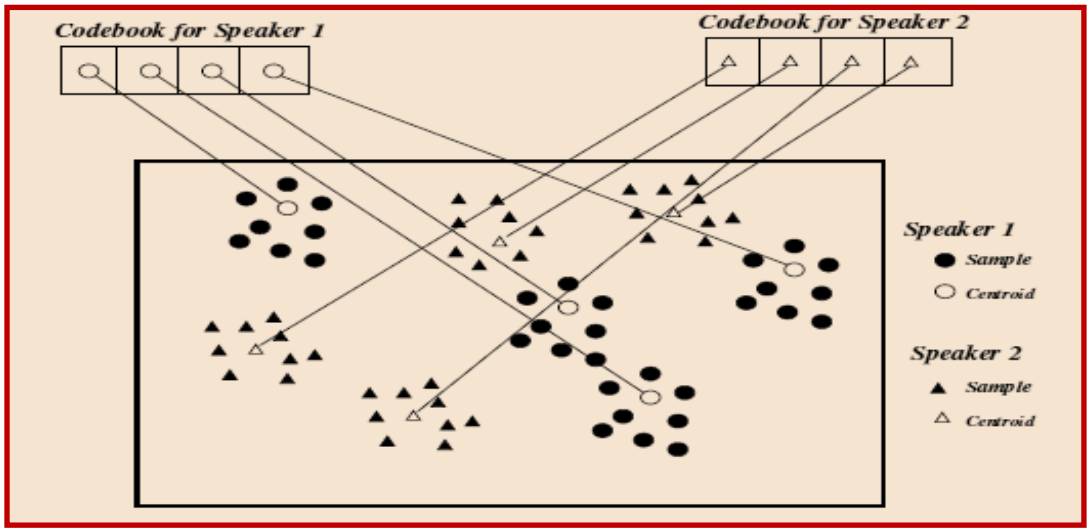

Figure 2.6: Vector quantization example

2.3.2.5 Support Vector Machine (SVM):

This is amongst the most powerful tools for in the field of pattern recognition techniques based on the application of a discriminative strategy. For data classification SVMs apply linear and nonlinear separating hyper-planes. But SVM cannot be applied to the tasks involving variable length data classification, because it can only classify fixed length data vectors. So there is a requirement of variable length data transformation into corresponding fixed length vectors before SVMs procedures starts. It is termed as a generalized linear classifier having maximum-margin fitting functions. The fitting function allows regularization that assists the classifier for better generalization. However, the classifier tends to overlook many of the features. For the purpose of controlling model complexity regular statistical and Neural Network models are applied by using a relatively small number of features (problem dimensionality/number of hidden units). Model complexity is controlled by controlling the VC dimensions of its model. This method is independent of dimensionality and makes use of spaces with greater dimensions spaces, that allows generation of huge non-linear features and later on applying adaptive feature selection in training phase [86]. In case VC dimensions are given, this approach can employ linear model by forwarding non-linearity to the features. For instance, as a regularized radial basis function classifier, a support vector machine may be deployed.

2.3.3 Artificial Intelligence Approach:

The Artificial Intelligence (AI) approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. AI takes advantage of the ideas pertaining to both methods. The main logic behind the AI approach to speech/speaker recognition is to compile and incorporate knowledge from a variety

of knowledge sources and then apply this on the task. The AI approach bids to convert the recognition procedure into mechanical as per the way a human being applies their intelligence and logic for visualizing, analyzing, and finally coming to a conclusion on the measured acoustic features. The information regarding linguistic, phonetic and spectrogram are used by Knowledge based approach. Recognition systems are developed:

- Using acoustic phonetic knowledge to develop classification rules for speech sounds.
- Again template based approaches are very nice in the designing a variety of speech recognition systems.
- Beyond these, a lot of linguistic and phonetic literature researchers have in sighted on understanding the human speech processing. Purely knowledge engineering design enforces expert's speech knowledge directly and explicitly into a recognition system. Such knowledge is primarily derived from extensive study and observation of spectrograms and later on incorporated applying rules or procedures. Another source of motivation for pure knowledge engineering also comes from the interest and research in expert systems.

Success of AI is hindered mainly by the following factors:

- The difficulty in quantifying expert knowledge.
- The integration of many levels of human knowledge such as phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. In other words- combining independent and asynchronous knowledge sources optimally remains a mystery.

Though indirectly, knowledge guides the design of the other models and algorithms such as template matching and stochastic modeling. Such application results in a marked distinction between knowledge and algorithms. The algorithms help users to solve problems, whereas the knowledge drives the algorithms towards improved performance. The successful strategy implementation in the speech recognition system is made possible by the contribution of this type of knowledge. Due to its use improvisation is seen in: The proper input representation selection decision, The speech unit meaning, and The algorithm design.

Thus, the AI approach to segmentation and labeling prefers to augment the generally used acoustic knowledge with phoneme knowledge, lexical knowledge, syntactic knowledge, semantic knowledge, and even pragmatic knowledge [52, 53].

2.3.3.1 Artificial Neural Networks

For employing artificial intelligence in speech recognition, several sphere of knowledge source acquisition is essential. The fundamental idea of artificial intelligence lies in:

- Learning i.e. capability of automatic knowledge acquisition and
- Adaptation i.e. adjusting with weight and error

Basic idea implemented by artificial intelligence(AI) approach is aiming to automate the process of speech/speaker recognition as per the similar manner a human being utilizes their intelligence in visualizing, analyzing, and characterizing vocal sound to absorb the conveys message through it by the speaker depending on some measured acoustic features. The neural network approach is a major approach through which these phenomenon are executed. ANN's primary focus has been in representing of knowledge and then integrating the knowledge sources. ANN (also termed as a

connectionist model) represents an in depth interconnection of computational elements that are simple but non-linear. The assumption is- there are *I* inputs which are to be summed up with weights, a threshold value and then nonlinearity and ultimately they are compressed to produce the output *O*. As the name suggests the term the neural network has been coined from the neurons of human brain (central nervous system). A critical factor in ANN is network topology which describes how the interconnection is established among simple computational elements. Three standard topologies generally applied are: Single/multilayer perceptrons, Recurrent networks (also known as Hopfield network), self –organizing networks (also called as Kohonen network)

The outputs at one layer (from one or more computational elements) forms the inputs the next layer in the single perceptron or MLP. On the other hand Hopfield network is recurrent by nature. Here the input to each element involves inputs along with outputs. Again, Kohonen network provides a clustering procedure. It gives a codebook of stable patterns in the input space. These patterns specify an arbitrary input vector using a small number of cluster representatives. Some worth mentioning benefits of ANN are listed below:

- Capable to instantly execute a huge volume of parallel computation. Because ANN is highly parallel yet simple structure. It consists of identical computational entities.
- Through the embedded data ANN distributes information to each element, and inherently ANN structure is immune to noise or defects within.
- Connection weights of the network is liberal, thus makes ANN suitable for real time application, hence results in improved performance.

- ANN facilitates nonlinear transformation among arbitrary inputs and outputs. Hence more efficient compared to available alternative tools in the field of physical nonlinearity implementations.

2.4 Speaker Recognition:

Human beings and computers recognise speech patterns in different ways (Ladefoged, 2001; Alexander et al., 2005). Humans use normally aural —hearing perception—, linguistic and other kinds of knowledge to identify others from their voices (Hollien, 2002; Rose, 2002); on the contrary, automatic speaker recognition systems rely on the use of acoustic properties extracted by computers (Rose, 2002).

Human vocal tracts in human body carry person wise unique physiological characteristics and anatomical size and shape. This makes every single speaker specific and different inherently. This factor serves as most significant in differentiation of speakers using their voice signals measurement. Process of recognizing a person by his/her voice is Speaker recognition. Voice belongs to cognitive biometric identity by virtue of the dissimilarity in anatomical structure of the speakers. Utterances generally contain certain message to convey to the listener, but along with it perceived information such as gender/identity/authenticity of the utterer, the environment etc are also carried, which are of great importance for research. In speech recognition/authentication, focus concentrates on extracting the message and overlooking auxiliary messages. On the contrary, in speaker recognition, the objective revolves arround collecting and analyzing speaker centric characterization and overlooking the conveyed message. SR has been a preferred choice of considerable number of researcher. As a result several propositions and developments are seen. If we look at it economically, Speech Recognition and

authentication system is now largely accepted in the market as it find application in large/mass level. Automatic speaker recognition/authentication technology is growing and will continue to grow for coming years as in many applications such as biometric personal identification, physical access control, computer data access control etc it has made its pathway.

Objective of Speaker recognition is materialized by means of two ways: *speaker identification* and *speaker verification or authentication*. The identification is an 1:N mapping procedure for the purpose of searching the best match for a speaker included in a closed set. It is performed by comparing the in hand voice against pre-recorded N voice templates in the database. on the other hand authentication normally gets its due in the role of a 'security guard' to maintain secured access of legitimate users and preventing the imposters or intruders. In forensic applications, normally authentication in conjunction of identification is established in a bid to manage a database of best matches after which authentication is performed a particular match is accepted or rejected.

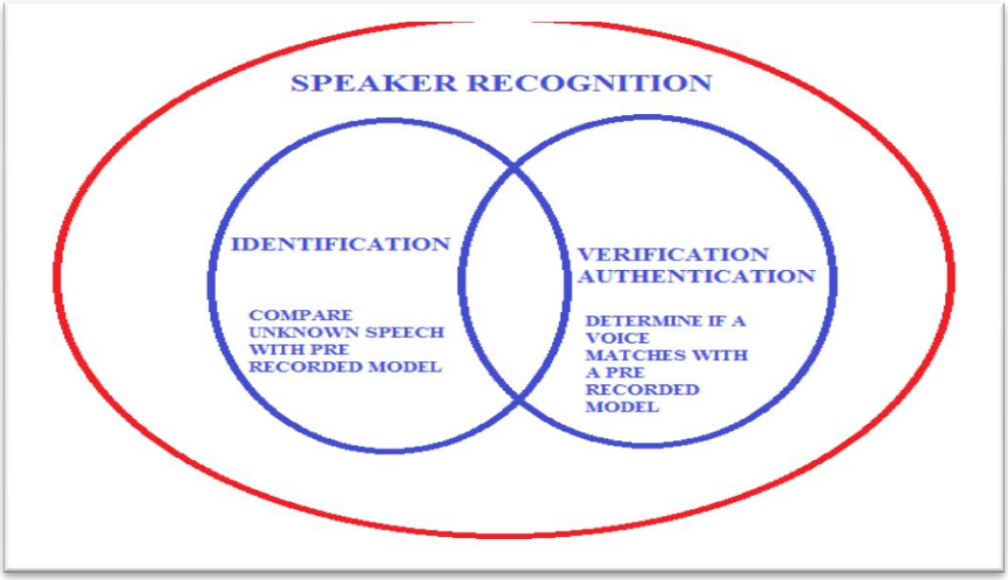

Figure 2.7: Identification, verification or authentication- 2 basic goals of SRS.

Some popular areas where identification is useful are: Real time intelligent feedback systems capable of generating user specific caller greetings, Automatic speaker-dependent audio indexing etc.

With a difference in approach, speaker authentication refers to a 1:1 mapping process to reach a conclusion whether to allow or deny the claimant voice. If result is positive identity claim of the speaker is authenticated otherwise declared imposter. In authentication procedure claimant's utterance checked for a matched with only one template in the pre existing database. Speaker authentication is a task to distinguish a voice of the claimant utterer known to the system from a potentially large group of voices unknown to the system. The speakers which are known to the system who claim their true identity are called *claimants*; speakers, either known or unknown to the system, who stand as other speakers are called *impostors*. There are two types of verification/authentication errors:

- ✓ False acceptances: when the system recognizes an impostor as a claimant;
- ✓ False rejections: when the system rejects a true claimant as an impostor.

Current applications such as computer log-in system; mobile banking, calling cards, and cellular-telephone fraud prevention substitute/supplement a memorized PIC (personal identification code) with speaker authentication. Authentication is suitable as an information rescue tool which can retrieve messages from a voice mailbox. Main objective of speech authentication is to characterize the uttered phonemes. On the other hand speaker is characterized in speaker authentication. From the point of view of utterance availability, speaker authentication process will be one of the following categories:

a) *Text dependent* (TD) and b) *Text independent (TI)*.

Primary condition involved in TD is that utterance must be closed to a finite set of pre recorded sentences. And the uttered text used for the purpose of training and testing the speaker authentication system must be within the identical set of voices. We can cite the example of an access-control authenticator where a claimant is allowed to use the personalized code again and again. So a authentication system carries the benefit of possessing the text to be spoken in advance. But this kind of authentication system also carries a risk of false acceptance of imposters if they try recorded voice of a legitimate client's phrase(s) and plays it back. Hence, to design secure system, the user enrollment text must be identical with authentication text. On the other side, absolutely unconstraint voices may be applicable in TI method. In TIS, utterances feed for training phase and during testing phase are comprehensibly independent and boundless. That brings upon the critically valuable flexibility in ground level deployments. As an instance we can cite the mail retrieval system where claimant speeches are unbound. Thus, in TIS, enrollment text most likely to differ from the testing text in most of the occasions. So we must ensure enrollment without client's awareness.

In between the limits of text dependence and independence there exists the *vocabulary-dependent system*, that constrains the speech to come from a limited vocabulary size, for example the digits (e.g., "one," "nine") from which test words or phrases (like "one-zero-nine") are formed. TIS provide a lot of flexibility compared to the TDS because pass phrases used by claimants can be altered on regular basis to facilitate thwarting an impostor with recorded utterance without retraining.

Certain speaker recognition/authentication tasks are accomplished by using models that extract and represent the best suitable information from the spectral sequence of speech wave. Mostly ASR systems depend on the spectral differences to discriminate informants. The main speaker-dependent information determined by the spectrum comes from vocal tract shape and size of the utterer. So, in this thesis we try to select a speaker model that in some sense captures the characteristic of vocal tract size and shapes of an informant's voice as manifested in the spectral features. Since success has been noticed in statistical pattern-recognition approaches for a wide variety of speech processing tasks, we amend a statistical formulation aiming such type of speaker model. In our work we treat the speaker as a random source producing feature vectors for statistical speaker model. There are a set of hidden states corresponding to characteristic of vocal tract configurations within these random speaker source. The spectral feature vectors are produced from a particular vocal-tract configuration, whichever state the vocal track was at that point of time. These states are considered to be hidden as we can't monitor the underlying states which produced the feature vector, but only the tangible spectral feature vectors.

Depending on the application, a SRS in general operate in two phases, namely *training* and *testing*. The system learns the voice characteristics of the speakers stored in the database of the system in the training phase. Feature vectors that are significant for the voice signal of the speaker are extracted and are applied later on in the formation of reference model applying the neural network training module. Same set of feature vectors are extracted from the test utterances (unknown) in the testing phase employing the same process. This testing is the actual recognition task. Now matching techniques provides the degree of their match. The level of match

after comparison to the pre specified threshold is considered for final decision that produce the ultimate conclusion whether to accept the test utterance or to reject it and send for further processing procedures. For pattern matching different statistical models are employed. Popular approaches are:

- ✓ Hidden Markov Models (HMMs), Gaussian Mixture Model (GMM) focuses on the underlying variations and temporal changes of the acoustic pattern.
- ✓ Dynamic Time Warping (DTW) focuses on measuring the similarity between two sequences that fluctuate in speed or time (even when the variation is non-linear e.g. when the speaking speed varies during the sequence).

Some of the popular SAS modeling techniques are: GMM (Gaussian Mixture Model), FVQ (Fuzzy Vector Quantization), LVQ (Learning Vector Quantization), SOM (Self-Organizing Map) etc. Success rate of these modeling paradigms also depends upon the conjugating clustering methodology. Modeling techniques are significant because a particular speaker is modeled applying one of these techniques through the feature vectors generated from the utterances they provide.

In the process of testing, each utterance speech features are considered to represent the test voice and matching is performed against every pre-determined models that exists in our database. Next step is to find the minimum distance (*Euclidean distance*) or maximum a posteriori probability. There exists some model in the database that resembles the test voice showing minimum distance or maximum a posteriori probability. This model will be the most likely speaker of the particular speech wave frame in consideration. Final decision is made as follows: ultimate speaker is one that receives highest count of frames in the test speech data. Main

contributors to the good authentication rate are: amount of data used in training and testing phase along with volume of Gaussian mixtures (size of codebook).

2.4.1 Speaker modeling using Direct Template Matching (DTM):

When the amount of available data is small, the number of feature vectors is also small. Since the number of feature vectors is insufficient, we can use direct template matching to get the speaker recognition rate [72]. In Direct Template Matching (DTM) technique to identify tentative speaker of the speech frame, in the entire identification phase, the feature vector to be verified (uttered by an unknown speaker) is compared with all the reference training feature vectors. The same process is looped until the last testing frame is reached. Speaker of the current voice is identified as the model which receives highest count of frames. Although procedure is easy for implementation, main objective- the recognition rate is poor. This is resulted from the inability to adapt with the large intra speaker and inter speaker variability. So is the need of different modeling techniques. The modeling technique can better the clustering or capture the improved distribution of the feature vectors as per the speaker information. Resultant speaker models carry the feature vectors of the same speakers while considering different sound units. This results in the dominance of speaker information over speech information.

2.4.2 Speaker Modeling through CVQ:

Vector Quantization (VQ) involves finding a subset of feature vectors named as Code vectors from the whole set, which can play as representative vectors for the test utterance. Aiming to model a particular speaker CVQ method clusters all the feature vectors in the feature space. Clusters are non-overlapping with crisp boundaries, and this is the reason behind its name. The codebook is the term used

for lookup table of code vectors. Applying KMC (k-means clustering) in conjugate to binary division codebooks is prepared in several sizes in training phase. *KMC results in* similar sets from input feature vectors into k-clusters that are non-overlapping. Individual codebooks are prepared against each and every speaker taking into account a particular size (optimum size 16) in the database. Ultimately target voice sample feature vectors are compared against the codebooks in the database and highest likely hood utterer is decided. However, research outcomes suggest better authentication rates for greater sizes like 32 and 64 also. However further increase in the codebook size (like 128) do not improve performance rather yields lower authentication rate. In totality CVQ authentication produces better results than DTM technique [175].

2.4.3 Speaker Modeling through FVQ:

This provides an alternative to just concluded CVQ and works upon fuzzy logic principle while clustering data feature vectors of the test voice sample [63, 208]. It is differentiated to other paradigms by virtue of the property that a feature vector can be assigned to more than one cluster. FVQ do clusters of all the feature vectors in the feature space into overlapping clusters with fuzzy boundaries In FVQ each feature vector is assigned to all the clusters by membership function, though degrees of association varies. As a result of this many to many association, more feature vectors signifies each cluster. This increases reliability. The codebooks of different sizes are constructed using binary split and fuzzy c-means clustering procedures during training (Bezdek & Harris 1978). Fuzzy c-means clustering associates grouping the input feature vectors into overlapping c-clusters. The nature of clustering depends solidly on the learning rate parameter. Hence it needs to be tuned

for better recognition rate. Finally like in SVQ feature vectors comparison is followed.

It gives a better recognition rate compared to the CVQ. The better recognition rate by FVQ shows that by increasing the number of elements for clustering, the recognition rate can be improvised. This is accomplished by correlating the same set of feature vectors to different clusters, obviously, by employing different membership functions. This improvement costs us of increased computational complexity of tuning the learning rate parameter. Even then, FVQ is preferable for small amount of data. On the similar lines we may walk around other VQ modeling techniques based on neural networks too.

2.4.4 Speaker Modeling Using SOM:

A neural network responds to VQ, but with unsupervised learning can be realized using Self Organizing Map (SOM). The approach for identifying the code vectors is by learning in an unsupervised way [86, 182]. The clustering is therefore affected by the actual distribution of feature vectors and therefore the modeling may be different. SOMs are a special class of neural networks which depends on competitive learning (Kohonen 1990). Hence, the performance of the SOM relies on the parameters such as neighborhoods (h), learning rate (η) and number of iterations. The recognition rate is quite high compared to other model even the feature vectors from limited data provide speaker information in the feature space. Further, each speaker has a unique distribution of feature vectors which is learnt by SOM.

2.4.5 Speaker Modeling through LVQ:

Kohonen in 1990s derived this supervised learning technique to globally optimize the codebooks. It uses class information for position optimization of code vectors

acquired by previous method, so as to improvised quality of the classifier decision regions. Selection of input vector is done randomly. The code vector is directed towards the input vector once the class labels of the input and the code feature vectors meets. Alternately, the code vector slips in opposite direction (far from the input vector). The improved authentication rate hints that employing supervised learning over unsupervised ones produces better authentication performance. Thus even in limited data restrictions also balanced application by LVQ is useful.

2.4.6 Speaker Modeling through GMM:

The Gaussian Mixture Model (GMM) is the most widely accepted probabilistic modeling technique in speaker recognition. The GMM wants sufficient data (at least one minute) to model the speaker well to yield high recognition rate (Reynolds & Rose 1995). In GMM system the distribution of feature vectors is modeled by the parameters like weight, mean and covariance (Reynolds & Rose 1995). The GMM yields the highest recognition rate compared to other past discussed models. The recognition rate of GMM-based system is better compared to CVQ, but poor compared to all other VQ modeling techniques. This means that the data may be too sparse to model by the Gaussian mixtures. To alleviate this problem to some extent the concept of Universal Background Model (UBM) can be used along with GMM.

2.5 Speaker identification and authentication:

As discussed earlier, speaker recognition is a biometric modality concerned with determining who the speaker is. It differs from speech recognition in that it does not attempt to draw any meaning from the message of what is spoken (NSTC Subcommittee of Biometrics, 2006). Instead, speaker recognition extracts and compares the unique physiological and behavioural features from the speaker's

voice in order to determine or verify the speaker's identity (Reynolds, 2002:4072-4075). There are two main classes of speaker recognition: speaker identification and speaker authentication (Jayanna and SR, 2009:181). These classes link directly to biometric identification and authentication.

Speaker authentication: Also referred to as speaker verification, is concerned with answering the question of "are you really whom you claim to be?" (Shen and Khanna, 1997:1436-1436), by either accepting or rejecting a user's claim of identity (Jayanna and SR, 2009:181). This is typically done by first performing a straightforward text-based one-to-many comparison between the username of the claimed identity and the usernames stored in the speaker database. If a username match is found, then a one-to-one comparison is performed between the voice token from the unknown speaker who is making the identity claim, and the voice token that of the username matched identity from in the speaker database. If this comparison is successful, then the identity claim will be accepted as valid. If the comparison is unsuccessful, the user's claim of identity is rejected (Doddington, 1985:1651-1664). SPERIA B will be able to perform speaker authentication by performing the process as outlined above. Speaker identification and authentication may be further classified into text-dependent and text-independent speaker recognition. Text-dependency in speaker recognition refers to whether or not the speaker recognition system requires the same text in the speech utterance during training and testing. Speaker recognition systems may be text-dependent or text-independent (Martsyshyn and Rashkevych, 2010:163-167), as discussed in the following sections. Text-dependent speaker recognition requires the same text to be spoken during the training and testing phases (Krishnamoorthy and Mahadeva Prasanna, 2009:729-754). The specific text is therefore known to the system and is

used as the "password", or passphrase, during testing, such as in the case of access control (Gold and Morgan, 2000:525-527).

It is important to understand that text-dependent speaker recognition should not be confused with speech recognition. Speech recognition is solely concerned with interpreting the message behind spoken speech, whereas speaker recognition is solely concerned with identifying and differentiating speakers based on physiological and behavioural/learned features present in their voice. If we attempt to apply pure speech recognition to spoken passphrase based access control, then the system will accept or reject an access attempt based on the correctness of the text in the passphrase, irrespective of who the speaker is. The problem emerges if an imposter successfully overhears this passphrase message from the authorised user; the impostor can make the identity claim of this user and simply repeat the passphrase in his or her natural voice, and the system will accept and authenticate the impostor. Speech recognition does not take into consideration physiological and behavioural features in the voice. In text-dependent speaker recognition, the passphrase message is not interpreted or understood by the system in any way. Instead, a detailed reference model is generated based on the unique physiological and behavioural characteristics present in the spoken passphrase during enrolment. This template may differ for the same individual if a different passphrase is spoken during testing, resulting in a speaker rejection.

2.5.1 Advantages of text-dependent system:

Text-dependent speech recognition offers a remedy to the problem as identified above in the case of speech recognition based passphrase access control. If the password text is successfully guessed or overheard by the imposter, this imposter

should not be able to authenticate as the authorised user even if the same passphrase message is spoken. The reason for this is that the reference model generated for the impostor will be different as the imposter has his or her own unique physiological and behavioural features. The use of speech recognition should not, however, be entirely discounted in speaker recognition. Some systems make use of text-prompted speaker recognition. Text-prompted speaker recognition would typically request that a user pronounce a sequence of words at random in order to verify that the user is human. Once the speech recognition component can verify the random sequence of words, the speaker recognition component then performs the authentication function (Gold and Morgan, 2000:525-527).

SPERIA (B) will employ a text-dependent speaker recognition method in both speaker identification and authentication functions.

2.5.2 Text-independent speaker recognition:

It does not require the same text during training or testing. In certain instances and applications, the text cannot be predicted during testing (Krishnamoorthy and Mahadeva Prasanna, 2009:729-754, Gold and Morgan, 2000:525-527). Text-independent speaker recognition is more convenient as users are able to speak freely to the system; however, it is more challenging problem to solve and generally requires longer utterances during training and testing phases (Liu, Huang & Zhang, 2006:1146-1149). In general, any spoken text should work in a speaker recognition system regardless of text-dependency (Gish and Schmidt, 1994:18-32), including voiced (vibrational) and unvoiced (frictional) speech. Since text-dependent speaker recognition can extract unique features associated with each syllable and phoneme,

it is generally realises a higher recognition performance compared to text-independent speaker recognition (Sadaoki, 1997:859-872).

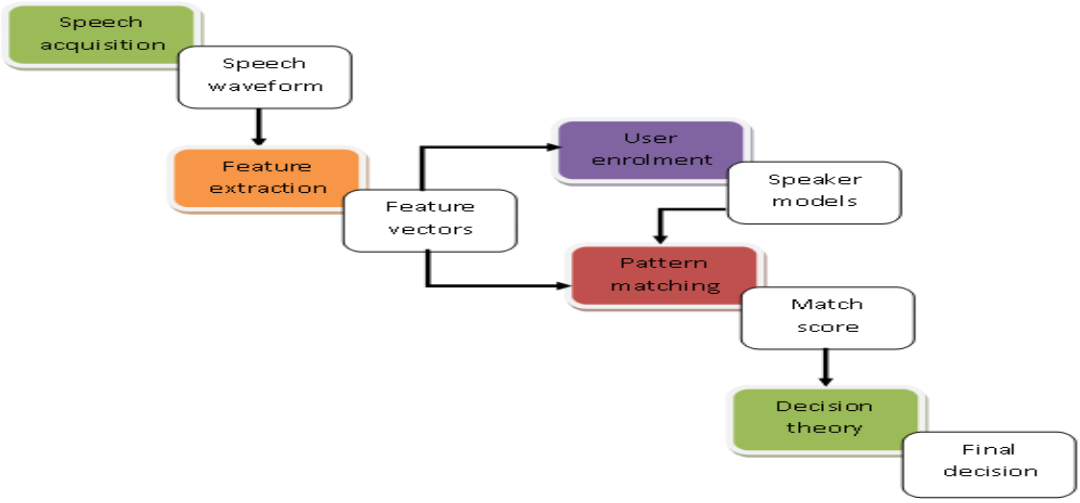

Figure 2.8: A generic speaker verification or authentication system

2.6 The Bodo Speaker Authentication System is Necessary:

Languages are the identity bearer of any community and very keenly connected to their lifestyle and socio economic culture. Existence of a wide variety of mutually exclusive diverse language families around the globe makes human civilization richer. Language of a particular community and their culture and environment are inseparably interlinked. Languages are closely connected to humans, culture, and environment. Its diversity has valuable contribution in the growth, development, and enrichment of human civilizations. Therefore, preserving linguistic diversity in the world is primary to keeping up a healthy cultural diversity which is necessary to mankind's prosperity [14, 17]. Worldwide statistics draws a disturbing picture of the extinction of several hundred languages. *In the last 500 years, almost half of the languages of the globe have already vanished and trend continues with other languages due to various factors.* Because of the tendency among the speakers favoring a more dominant/popular language, from which they gets more economic or cultural exposure, starts avoiding their own language. In the rapidly changing

political, economic, technological, and socio-cultural fronts worldwide (termed as globalization), a phenomenon like language globalization also taking place threatening the very existence of the languages (many such languages) having relatively less number of speakers belonging to smaller communities. The term Globalization represents a unified society for the people world over to live as a single society and progress uniformly and united. As per Linguists expert's estimation extinction rate is so rapid that up to ninety percent of existing languages may vanish within a few generations. Researcher must device some technology so that this dangerous trend can be averted with utmost sincerity. Natural language processing technology can have immense contribution in this regard to save/preserve these languages. Systems like ASR, SI/SA developed for such languages will definitely be a good beginning in this front. As off now out of 7000 languages around the globe, only a few of them have their recognition / authentication systems developed so far. Moreover those with a large population (with a high economic value and political importance) systems are developed. Whatsoever, change in situation sometimes leads to shift in focus for language development all of a sudden. Such an example is Arabic language. The deadly terrorist attacks on the World Trade Center in New York City and the Pentagon in Washington, D.C., resulted in war in Afghanistan. After this Arabic language became a hot favorite for Global Autonomous Language Exploitation (GALE) project sponsored by the Defense Advanced Research Agency (DARPA) in the United States, although for helping the US fighters. Similarly after the start of the second Gulf War and the invasion of Iraq, Iraqi language became of interest to the American military, and DARPA started a project in Transtac. The so called *killer* language tag mainly goes to English. As an instance we can cite the example of Gaelic and Irish languages. Despite having a

large speaker community they are facing grave danger of being replaced by English because of the preference of the speakers. An inclination for English has grown worldwide in recent years. The BODO language, a dominant tribal language of India with its inherent tonal melody, is too not immune to such adverse effects and may move towards extinction. Very little efforts have been put in Bodo Speaker Authentication till date. So this research work is pioneer in nature. The three layered approach applied in the feature selection in this work is unique.

Against all the odds, an automatic speaker recognition/authentication system for BODO language (SPERIA-B) will be helpful in order to keep the language alive and maintain diversity and its purity. As this is the beginning, obviously our research work involves a small number of vocabularies that are generally used in the small recognition/authentication system. There are huge diversity of phonetic content in the languages spoken in India. It holds good for even BODO language that has evolved in the race of time beginning in original roots after several editions. A great influence of other local dialects of Assam and North-Eastern part of India has contributed to this process as well. So a strong need is felt for extensive studies for design of speech Identification, verification/ authentication system for *BODO language exclusively*.

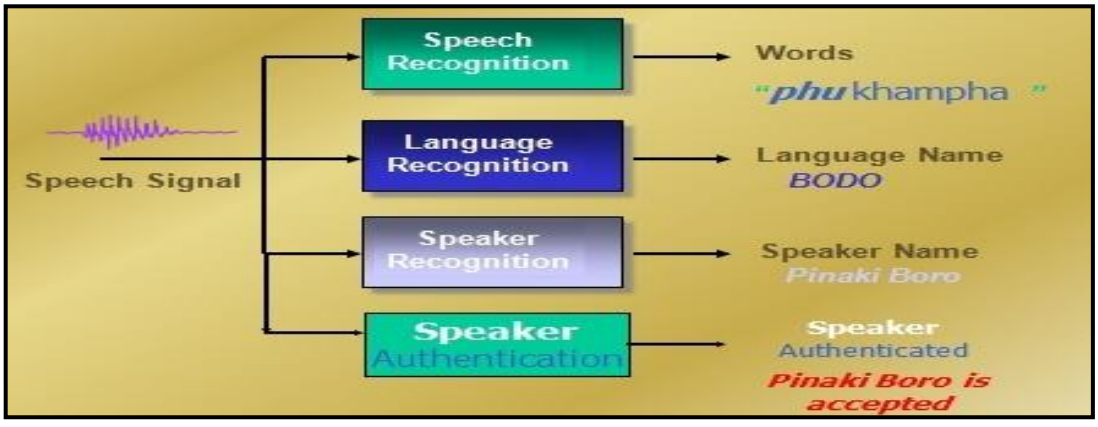

Figure 2.9: Authentication approach used in this research work

2.6.1 Evaluation of a biometric verification system:

After having computed a match score of similarity between the input user and the corresponding template stored in the database, a decision is taken whether the user must be accepted or rejected by the system. However, such decision can be either correct or not correct. If the decision is incorrect, two different types of error can occur (Bimbot et al., 2004):

- False rejection (or non detection): the system rejects a valid identity claim.
- False acceptance (or false alarm): the system accepts an identity claim from an impostor.

Both types of errors give rise to two types of error rates, which are commonly used to measure the performance of a system:

- False rejection rate (FRR): percentage of incorrectly rejected clients.
- False acceptance rate (FAR): percentage of incorrectly accepted impostors.

Either of the two types of errors can be reduced at the expense of an increase in the other, so that the trade-off between FRR and FAR depends on a decision threshold. In a real-world system, which is usually not perfect, FRR and FAR intersect at a certain point (depicted in adjoining Figure A and B). The value of FRR and FAR at this point is known as the Equal Error Rate (EER). If the threshold is set to a low value, the system tends to accept most of the identity claims, giving few false rejection errors but many false acceptances. On the contrary, with a high threshold the system tends to reject most of the identity claims, giving rise to few false acceptance errors and a lot of false rejections.

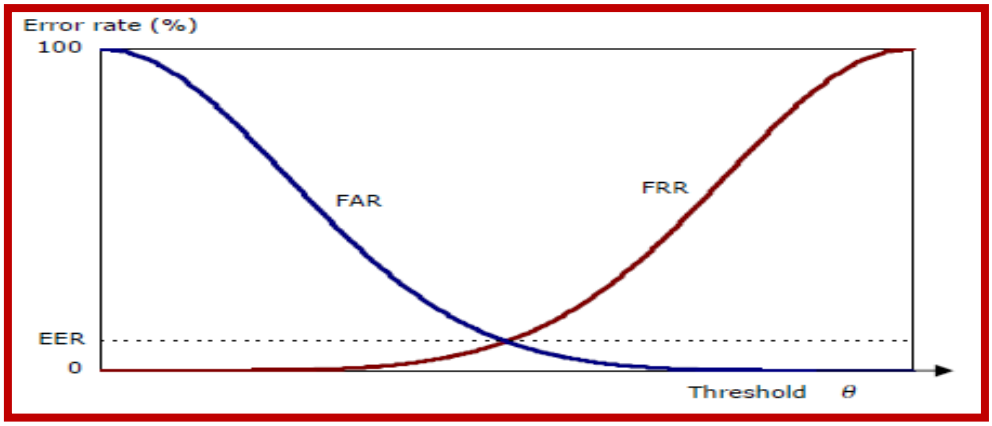

Figure 2.10 [A]: FRR & FAR as a function of a threshold ɵ. Intersection point determines the EER.

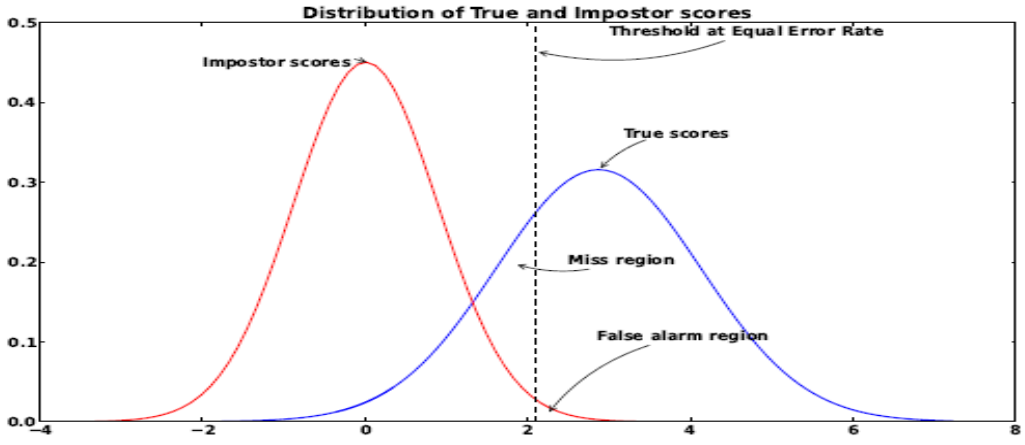

Figure 2.10 [B]: Illustration of Equal Error Rate for a given set of true and impostor scores

Therefore, when designing a biometric verification system, the decision threshold must be adjusted so that both errors are as low as possible, or one of the errors must be always below a certain threshold, if this property is required by a specific application. The trade-off between the two types of error rates is usually depicted in different ways. Two of the most common representations are the ROC and the DET curves. The Receiver Operating Characteristic (ROC) curve plots the FRR versus the FAR (Bimbot et al., 2004). This curve is monotonous and decreasing, and the better the system is, the closer to the origin the curve will be. Another representation

of the ROC curve is used sometimes by plotting the correct detection rate (instead of FRR) versus the false alarms (Duda et al., 2001). It is also common to plot the error curve on a normal deviate scale. In this case, the curve is known as the Detection Error Trade-offs (DET) curve (Martin et al., 1997; Bimbot et al. 2004).

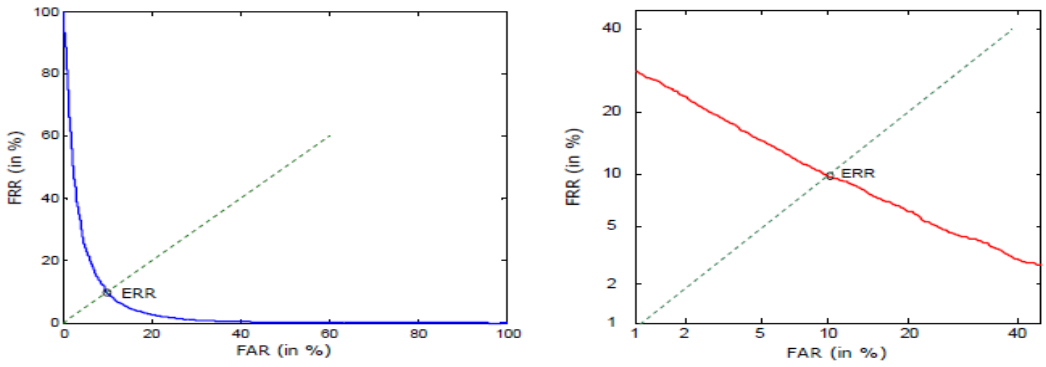

Figure 2.11: Examples of a ROC curve (a) and the corresponding DET curve (b).

Chapter Summary: In this chapter, we have focussed on providing a solid background into the fundamentals and development of speaker recognition and Authentication. We started the chapter with an overview of speech-enabled human computer interaction, where we discussed and differentiated speech and speaker recognition, followed by a history into speaker recognition. We then discussed speaker identification and authentication. We then looked at text-dependent and text-independent speaker recognition. Finally the performance evaluation of SA system is discussed.