ISSN:0975-9646

Synthesis of Speech & Speaker Recognition Using Data Driven Approach

Leela Kumari , Rahul Guha , Geetika Gera

Computer Science Department,
Dr. Radhakrishnan Institute Of Technology,
Jaipur, India

***Abstract*— As we can discuss about your Personal Computer then it would provide a comfortable and natural form of communication. This will reduce the typing amount, and it would be effortless, and allow you to move away from the terminal or screen. You will not have to be in the line of sight of the terminal. It will help you in many cases to tell you that who was speaking.**

If i want to use voice as a new medium on a computer workstation, it is natural explore how speech recognition can contribute to such an environment. Here we will review the state of speech and speaker recognition, focusing on current technology applied to personal workstations. The objective of this paper is to provide explanation of the speech and speaker recognition data input of the user. An algorithm which efficiently determines the optimum coordination of H.M.M has been successfully designed with the help of MATLAB.

***Keywords*—Automation system for pattern matching, MATLAB coding speaker identification, Speech processing.**

I. INTRODUCTION

The capability of Speech recognition, to find out the words spoken, and recognize the speaker, the ability to identify who is saying them, will become commonplace applications of speech processing technology. The forms of the recognition of speech are available for personal workstations. Now days the interest in speech recognition has increased more, and improving its performance. Speech recognition has proved very useful for certain applications, such as telephone voice-response systems for selecting services or information, digit recognition for cellular phones, and data entry while walking around a railway yard or clambering over a jet engine during an inspection. Nonetheless, comfortable and natural communication in a general setting (no constraints on what you can say and how you say it) is beyond us for now, posing a problem too difficult to solve. Fortunately, we can simplify the problem to allow the creation of applications like the examples just mentioned. Some of these simplifying constraints are discussed in the next section. Speaker recognition is related to work on speech recognition. Instead of determining what was said, you determine who said it. Deciding whether or not a particular speaker produced the utterance is called *verification,* and choosing a person's identity from a set of known speakers is called *identification.* The most general form of speaker recognition (text-independent) is still not very accurate for large speaker populations, but if you constrain the words spoken by the user (text-dependent) and do not allow the speech quality to vary too wildly, then it too can be done on a workstation. Speech processing comes as a front end to a growing number of language processing i.e., it is a diverse field with many applications. In this Paper, I have focus on phonemes and allophones in order to try to provide a deeper insight into the kinds of issue involved in the processing of speech.

II. REVIEW OF LITERATURE

The development of speech recognition systems began as early as the 1960s with exploration into voiceprint analysis, where characteristics of an individual's voice were thought to be able to characterize the uniqueness of an individual much like a fingerprint. The early systems had many flaws and research ensued to derive a more reliable method of predicting the correlation between two sets of speech utterances. Speaker identification research continues today under the realm of the field of digital signal processing where many advances have taken place in recent years. The first speech recognition system having creditable performance was built in 1952 at Bell labs using acoustic features to recognize digits spoken by single speaker. But these techniques and results could not be extrapolated towards larger and more sophisticated systems. The research has been carried out with acoustic phonetic approach in mid 1970's. The advanced research project agencies support of speech understanding research has lead to significantly increased level of activity in this area since 1971. Several isolated and connected speech recognition systems have been developed and demonstrated.. For speech signal processing fast Fourier transform, cepstral analysis, and linear predictive coding were started to be used. New techniques for pattern matching like DTW (dynamic time warping) and HMM were invented. Linde, Buzo and Gray presented an efficient and intuitive algorithm for the design of vector quantizers. Gaussian mixture techniques were also used to provide a possible useful extension to the popular EM (expectation maximization) algorithm [6]. To improve generalization capabilities of HMMs in many practical classification problems large margin HMM based classifiers are used.

III. COMMUNICATION-THEORETIC FRAMEWORK

A Speech communication system involves sensory as well as cognitive behaviors. Traditionally, the speech communication chain comprises four stages: detection of acoustic-phonetic cues to form words, syntactic and grammatical analysis for parsing of sentences and for error correction, semantic determination and disambiguation, and pragmatic inference with additional prosodic cues and

interpretation of message intent. The ultimate machine that can converse with a human would need all the knowledge to perform those four stages of the speech communication chain. It involves understanding of the context (the subject domain as well as the mood and the ambient) of the conversation. No attempt has yet reached such a level of complexity.

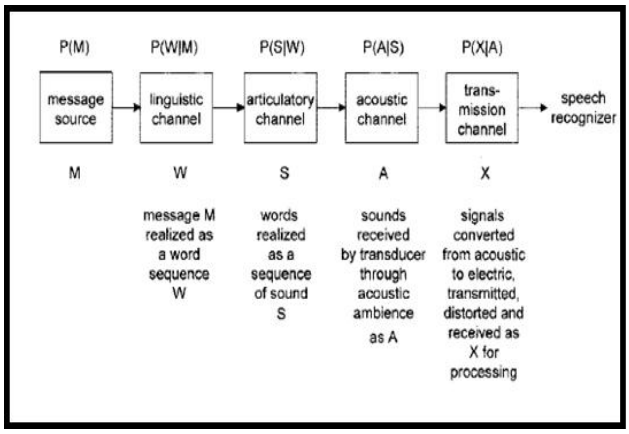

Fig. 3.1 - Communication-theoretic formulation of the speech production chain.

Following this limited goal of human– machine communication, a concrete and yet convenient way to describe the speech generation/production chain is shown in Fig. 1.1, which depicts the basis of a communication-theoretic approach to automatic speech recognition and understanding. In this formulation, a message source decides to convey an intended message M, which is realized as a word sequence W through a linguistic channel, specified by more likely than others. For example, Fig. 1.2 shows partially a finite state network of numerous expressions, all leading to the same semantic message: *the user needs information about flights that leave Dallas*. The word sequence W then gets realized, through the articulator channel,

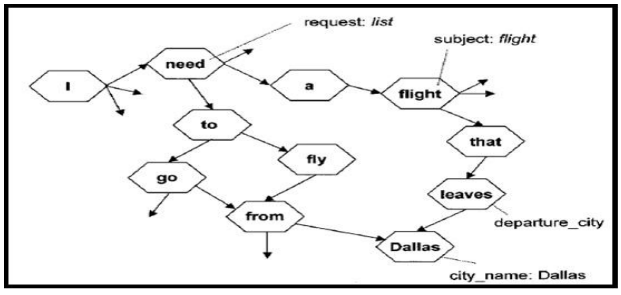

Fig.3.2- Finite state network

We label this process “acoustic channel.” The acoustic ambient in various rooms can be quite different. Fig. 1.4 shows the power spectrum of a typical acoustic background noise in a personal office with a computer running. The acoustic channel is characterized by a probabilistic model P (A|S).

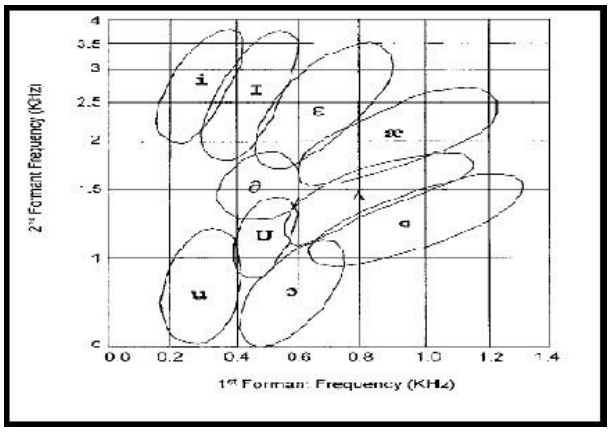

Fig. 3.3 - Distribution of vowels in F1-F2 plane

IV. SPEECH SIGNAL MODELLING

A. Acoustic Modelling

In this subsystem, the connection between the acoustic information and phonetics is established. Speech unit is mapped to its acoustic counterpart using temporal models as speech is a temporal signal. The models used for this purpose are hidden Markov model (HMM), Artificial Neural Network, Dynamic Bayesian Network (DBN) Acoustic modeling aims at finding the probabilistic behavior of the given data, expressed in the form of P(X|_) . [Here, we use P(X|_) in lieu of P_(X|_) without ambiguity because the class label is implied in the context.] This is often referred to as probability distribution estimation; i.e., finding the parameter in a certain optimal sense to define the distribution P (X|_).

1) Probability Distribution for Speech: The statistical method, as discussed in the previous sections, requires that a proper, usually parametric, distribution form for the observations be chosen in order to implement the MAP decision. Using the task of isolated-word speech recognition as an example, we have to determine the distribution form for the speech utterance of each word before we employ an estimation method to find the values of the parameters.

2) Speech Model: Based on the above characterization of the speech signal, a reasonable speech model or distribution should have the following three components. First, at an interval on the order of 10 ms, short time measurements are to be made along the pertinent speech dimensions that best carry the relevant information for linguistic distinction. These dimensions determine the observation space in which the distribution is to be defined. This is accomplished in signal analysis and the choice of representation.

B. Feature Selection & Measure

For appling mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. The selection of appropriate features and methods to estimate (extract or measure) them are known as feature selection and feature extraction, respectively. Traditionally, pattern-recognition paradigms are divided into three components: feature extraction and selection, pattern matching, and classification. Although this division is convenient from the perspective of designing system components, these components are not independent.

C. Pattern Matching

The pattern-matching task of speaker verification involves computing a match score, which is a measure of the similarity between the input feature vectors and some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is generated and stored (possibly on an encrypted smart card). Then, to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user. There are two types of models: stochastic models and template models. In stochastic models, the pattern matching is probabilistic and results in a measure of the likelihood, or conditional probability, of the observation given the model. For template models, the pattern matching is deterministic.

1) Template Model: The simplest template model consists of a single template $\overline{\mathbf{X}}$, which is the model for a frame of speech. The match score between the template $\overline{\mathbf{X}}$ for the claimed speaker and an input feature vector $\mathbf{X}_i$ from the unknown user is given by $d(\mathbf{X}_i, \overline{\mathbf{X}})$ The model for the claimed speaker could be the centroid (mean) of a set of N training vectors:

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i \qquad \text{…….. (8)}$$

Many different distance measures between the vectors $\mathbf{x}i$ and $\mathbf{x}$ can be expressed as:

$$d(\mathbf{x}_i, \overline{\mathbf{x}}) = (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\mathbf{W}(\mathbf{x}_i - \overline{\mathbf{x}}) \qquad \text{…..} (9)$$

where $\mathbf{W}$ is a weighting matrix. If $\mathbf{W}$ is an identity matrix, the distance is *Euclidean;* if $\mathbf{W}$ is the inverse covariance matrix corresponding to mean $\mathbf{x}$, then this is the *Mahalanobis distance.*

2) Dynamic Time Wrapping: The most popular method to compensate for speaking-rate variability in template based systems is known as DTW. A text-dependent template model is a sequence of templates $(\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_N)$ *N* that must be matched to an *input sequence* $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ .In general, *N* is not equal to *M* because of timing inconsistencies in human speech. The asymmetric match score *z* is given by

$$z = \sum_{i=1}^{M} d(\mathbf{x}_i, \overline{\mathbf{x}}_{j(i)}) \qquad \text{…….(10)}$$

where the template indices *j(i)* are typically given by a DTW algorithm. Given reference and input signals, the DTW algorithm does a constrained, piecewise linear mapping of one (or both) time axis(es) to align the two signals while minimizing *z*. At the end of the time warping, the accumulated distance is the basis of the match score.

3) HMM Model: A stochastic model that is very popular for modeling sequences is the HMM. In conventional Markov models, each state corresponds to a deterministically observable event; thus, the output of such sources in any given state is not random and lacks the flexibility needed here. In an HMM, the observations are a probabilistic function of the state; i.e., the model is a doubly embedded stochastic process where the underlying stochastic process is not directly observable (it is hidden). The HMM can only be viewed through another set of stochastic processes that produce the sequence of observations . The HMM is a finite-state machine, where a pdf (or feature vector stochastic model) $p(\mathbf{x}|s_i)$ is associated with each state *si* (the main underlying model). The states are connected by a transition network, where the state transition probabilities are a_{ij}

$p(s_i \mid s_j.)$ For example, a hypothetical three-state HMM is illustrated in Figure 4.1.

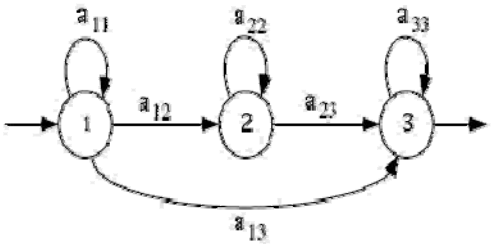

Fig.4.1- HMM Model

The probability that a sequence of speech frames was generated by this model can be found by Baum-Welch decoding . This likelihood is the score for L frames of input speech given the model:

$$p(\mathbf{x}(1;L)|\text{model}) = \sum_{\substack{\text{all state}\\ \text{sequences}}}\prod_{i=1}^{L} p(\mathbf{x}_i|s_i)\,p(s_i|s_{i-1}) \qquad \text{……} 12)$$

4) Spoken Dialog System: Having decoded the speech signal into a sequence of words, or a hypothesized sequence of words, a traditional speech understanding system employs a sentence parser to cast the word sequence into a structure to allow syntax verification and inference of meaning. The coupling between parsing and understanding is,

however, not a particularly tight one because most parsing algorithms focus on the linguistic structure first, rather than understanding. At the present time, the goal of automatic understanding of speech is limited to determining the required action based on the speech input. Telephone call routing, which connects a call to a proper destination based on the spoken query, is one such example. Most of these queries involve a single action to be taken by the machine. Another simple speech understanding task that has been attempted is DARPA's Air Travel Information System (ATIS). In the system, the user talks to the machine to obtain flight information such as "I would like to leave New York for San Francisco on November first, please list the available flights"; "How much does the flight cost from Dallas to Detroit?" In this task, the action to be taken involves the need to cope with the language structure in order to decide which information is to be provided to the user. For example: Any blank in the template or ambiguity in the input would induce a question from the system for further interaction. Clarification and follow-up questions can be initiated based on the information in the template as well as a temporary cache that records intermediate answers.

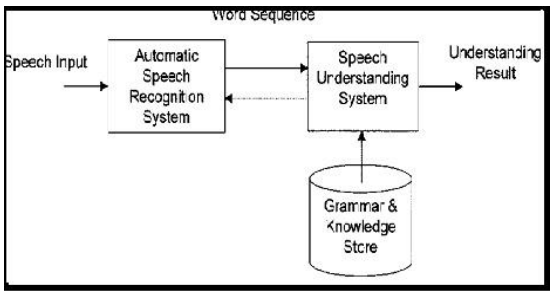

Fig.4.2- Block Diagram of an understanding system based on speech to word conversion.

D. Database and Generalization

The data-collection mechanism has to be attended with care. For example, a system deployed for digital cellular phone users needs to take into account the speech coder characteristics. But, an improperly designed ant aliasing filter for analog-to-digital conversion at the front end of the system should not be considered part of the adverse effect or source of variability and needs to be corrected before data collection can begin. The confusion between recorded diversity in operating conditions and the unwanted interference or adverse effects due to misuse of equipment often exists. An unwanted interference will cause detriment in the training result, but a true coverage of the diverse operating condition (e.g., real ambient noise) is crucial in guaranteeing a satisfactory performance. Thorough understanding and examination of the signal is very important.

E. Fourier Transform

Digital sound synthesis is the reverse operation. It is carried out using the digital to- analog converter (DAC) of the sound card that maps digital codes onto analog signals. Sounds can be amplified and made audible by a loudspeaker. Sound file formats might be different for different operating systems or software. Sounds can be stored and played back using, for example, .wav or .au files. Sound recordings can be digitized in other formats and compressed to save storage. A discrete Fourier transform and its computerized optimization, the fast Fourier transform (FFT), are methods to decompose digital time signals into their corresponding sinusoids and hence to map a signal onto its composing frequencies. The frequency set of a signal is also called the Fourier spectrum. Fig. 1.8 shows examples of Fourier transforms for some functions.

V. IMPLEMENTATION

During the first experiment a program has been written in MATLAB to verify the characteristic of speech and speaker recognition. During the program first we enter the name to be verified, which should be done within two second. There is a choice of entering the data voice again also, by pressing 1 else the program will verify the voice with the stored data. During the first program we have speech the correct word, and the output is as:

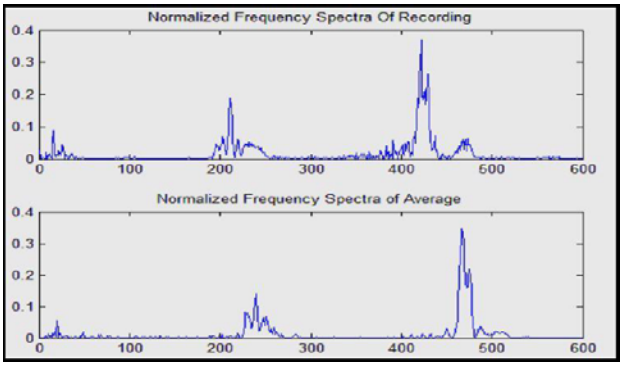

Fig.5.1- Voice Match with Pre-recorded Voice.

During the second experiment a program has been written in MATLAB to verify the characteristic of speech and speaker recognition. Flow of the program is same as in the first program.During the second program we have speech the wrong word, and the output is as:

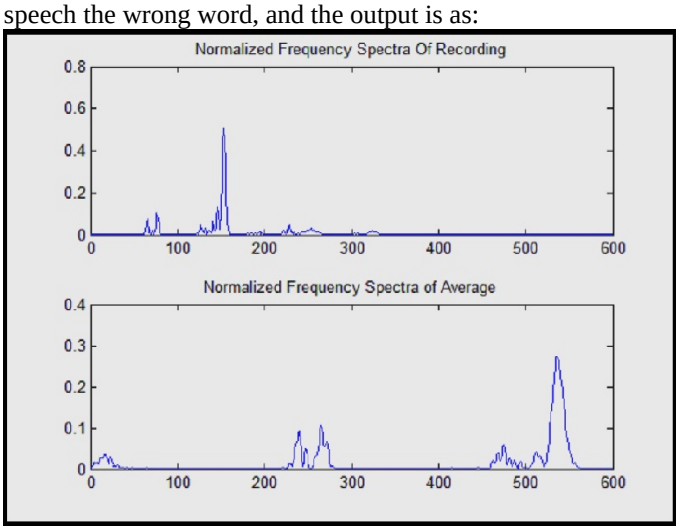

Fig. 7.2- when voice do not match with pre recorded voice

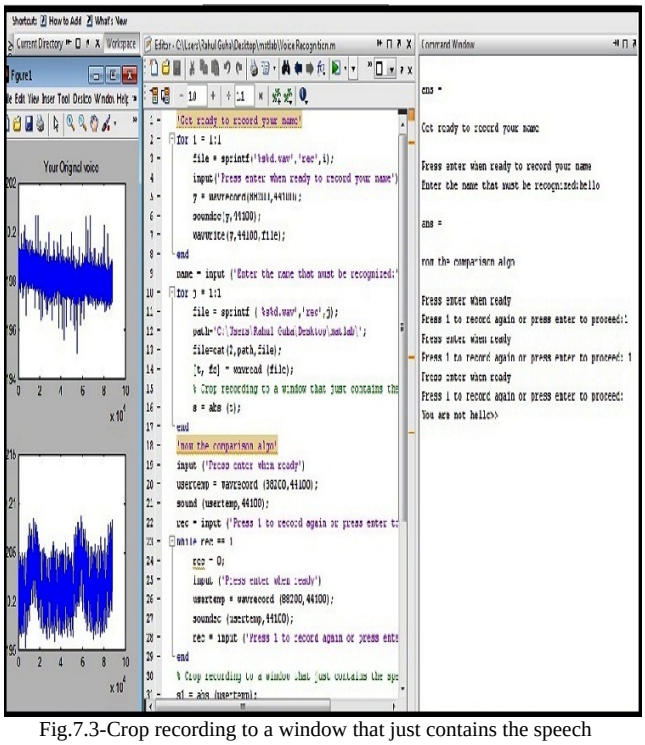

Fig.7.3-Crop recording to a window that just contains the speech

VI. FUTURISTIC ISSUES

Over three decades of research in spoken language processing have produced remarkable advances in automatic speech recognition and understanding that helps us take a big step toward natural human–machine communication. Signal-processing techniques led to a better understanding of speech characteristics, providing deep insights into acoustic-phonetic properties of a language. The introduction of a statistical framework not only makes the problem of automatic recognition of speech tractable but also paves the road to practical engineering system designs. It was found that a particular probabilistic measure, the HMM, provides a speech modeling formalism that is powerful and yet easy to implement. Coupled with a finite state representation of a language, hidden Markov modeling has become the underpinning of most of today's speech-recognition and understanding systems under deployment. To accomplish the ultimate goal of a machine that can communicate with people, however, a number of research issues are awaiting further study. Such a communicating machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions and have an efficient way of representing, storing, and retrieving "knowledge" required in a natural conversation. With the current enthusiasm in research advances, we are optimistic that the Holy Grail of natural human– machine communication will soon be within our technological reach.

VII. CONCLUSION

The objective of this thesis is to provide some explanation of the speech and speaker recognition data input of the user. An algorithm which efficiently determines the optimum coordination of H.M.M has been successfully designed with the help of MATLAB. Authentication of the user can be determined by the threshold value being set by the standard variance.

ACKNOWLEDGMENT

Our thanks to the experts who have contributed towards development of the paper. We also want to thank Dr.Saroj Hiranwal for her contribution and preferable guidance.

REFERENCES

[1] K. Lee, Automatic, Speech Recognition: the development to f the Sphinx System, Kluwer Academic Publishers, Norwell, Mass. 1989

[2] L. R. Bahl et al., "Speech Recognition of a Natural Text Read as Isolated Words," Proc. IEEE IntI Conf. Acoustics, Speech, and Signal Processing, April 1981, pp. 1,168-1,17 1.

[3] D.O. Kimbal et al., "Recognition Performance and Grammatical Constraints," Proc. DARPA Speech Recognition Workshop, Feb. 1986, pp. 53-59.

[4] D. O'Shaughnessy, Speech Communication: Human and Machine, Addison- Wesley, Reading, Mass., 1987.

[5] M.A. Franzini, M.J. Witbrock, and K.-F. Lee, "Speaker-Independent Recognition of Connected Utterances Using Recurrent and Non recurrent Neural Networks," Proc. Int'l Joint Conf. Neural Networks, V01.2, Washington, DC, June 1989R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] J. Mariani, "Recent Advances in Speech Processing," Proc. IEEE Intl Conf. Acoustics, Speech, and Signal Processing, Glasgow, Scotland. May 1989, pp. 429-440.

[7] M.-W. Fung et al., "Improved Speaker Adaptation Using Text-Dependent Spectral Mappings," Proc. IEEE Int'l Conf. Acoustics, Speech. and Signal Processing, New York City, 1988, pp. 131-134.

[8] D.B. Paul, "The Lincoln Robust Continuous Speech Recognizer," Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, Glasgow, Scotland, 1989, pp. 449- 452.

[9] H. Murveit and M. Weintraub, "1,000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models," Proc.