ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

A Review on Free Model Speech Recognition System Using MFCC Model

Bhavna, Dr.Dinesh Kumar

M.Tech Student, Dept. of CSE, SRCEM, Palwal, MD University, Haryana, India

Professor & HOD, Dept. of CSE, SRCEM, Palwal, MD University, Haryana, India

ABSTRACT: This paper takes a tour of speech recognition system which includes it's basic working, expectations of user from it, techniques involved in speech recognition and difficulties faced during the speech recognition process. Despite of more than 60 years of research in this field, variations in context, environment and speakers is still a major challenge faced by it.

KEYWORDS: Speech Recognition system, Analysis, Feature Extraction, Modeling techniques, MFCC, HMM & N-Gram.

I.INTRODUCTION

Among human beings, speech is considered to be the principal mode of communication as it is natural as well as efficient way of exchanging one's views, thoughts and information with the other(s).

A Definition

Speech Recognition is the process of converting speech signal (fed as input to the speech recognizer) to a sequence of words, using a computer program. It is also well known as Automatic Speech Recognition (ASR).

B Types of Speech [1] [2]

Speech recognition systems vary on basis of the way they accept utterances as input.

1) Isolated Words : These recognizers require isolated utterances, i.e. the speaker must wait in-between utterances so that the recognizer can do the required processing during this interval.

2) Connected Words : These recognizers require connected utterances as inputs, i.e. the speaker is allowed to speak separate utterances together with minimum pause between them.

3) Continuous Speech : Consider to be the most difficult recognizers to create, it is like a computer dictation to the speaker.

4) Spontaneous Speech : Here the natural speech of the speaker is fed as input to the computer.

C Expectations from a speech recognizer [3]

The ideal goal that a speech recognizer must achieve is : "Understand instantly and correctly everything the user wants you to hear, and nothing he/she doesn't".

However, it is impractical to expect this from the recognizer and thus the following attainable goals of speech recognizer are :

- To work effectively in noise as well as no-noise (or complete silence) environments.
- To work effectively for the widest possible range of heterogeneous inputs.
- To respond instantly to operator speech, to eliminate both cost and user frustration caused by delays.

II. SPEECH RECOGNITION SYSTEM

A Basic Working [1] [3]

Speech recognizers make it possible for the computers to understand human speech. Speech recognizers categorize vocabulary as: active vocabulary and vocabulary. Active vocabulary denotes list of words the user can be expected to say at any instant. While vocabulary denotes list of words the user may speak while working with the application. Speech recognizer loads a set of sound reference patterns that the application expects user to say. The recognizer classifies the unknown sound and reports the best possible match with reference patterns.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

The probability of occurrence of a word within given acoustic observations i.e. P(W/A), is given as follows (using Baye's rule format) :

$$P(W/A) = \frac{P(A/W)\,P(W)}{P(A)}$$

where P(A/W) is called the acoustic model that estimates probability of a sequence of acoustic observations on word string W P(W) is the language model that describes probability of a sequence of words.

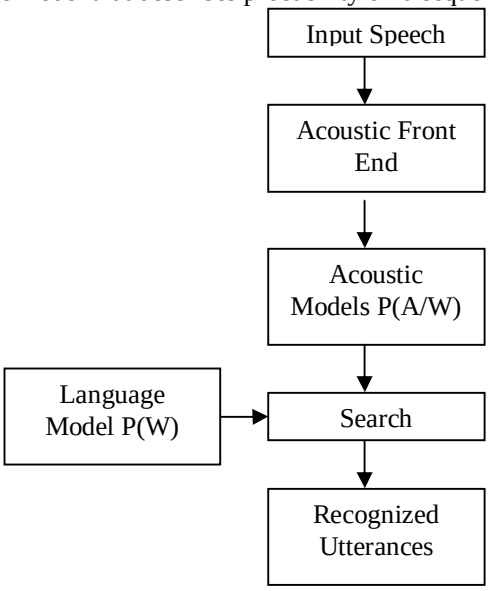

Fig. 1 Basic Model of Speech Recognition

B Speech Recognition Techniques [2][3]

The main goal of speech recognizer is to analyze, extract characteristics and recognize spoken information. Four stages in this process are :

1) Analysis : i) Speech Analysis Technique – Speaker is identified based on his vocal tract, excitation source and background feature. *ii) Segmentation Analysis* – The speaker is recognized based on frame size and shift in range of 10-30 ms to extract vocal tract. *iii) Sub Segmented Analysis* – Speech is analyzed based on frame size and shift in range of 3-5 ms to extract characteristics of excitation state. *iv) Supra Segmental Analysis* – Using frame size, characteristic of speaker due to behavior character is obtained.

2) Feature Extraction : As the number of given inputs increase, the number of training and test vector needed for classification also increase. So feature extraction step is quiet essential with it's focus on the following features –

i) Spectral features like band energies, formats, spectrum and cepstral coefficient i.e specific to vocal tract. ii) Variation in pitch or excitation source. iii) Behavior features like duration, information energy. Feature extraction basically revolves around two steps : training and testing. During training phase, the system is familiarized with the voice characteristics of the registering speaker by building reference models that are extracted from training utterances.

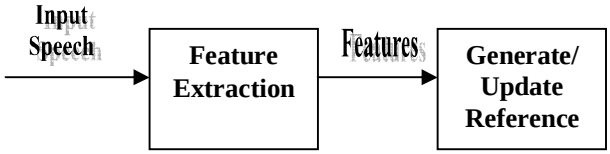

Fig : Training phase [7]

During training phase, from the test utterances similar feature vectors are extracted and matched with the reference model.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

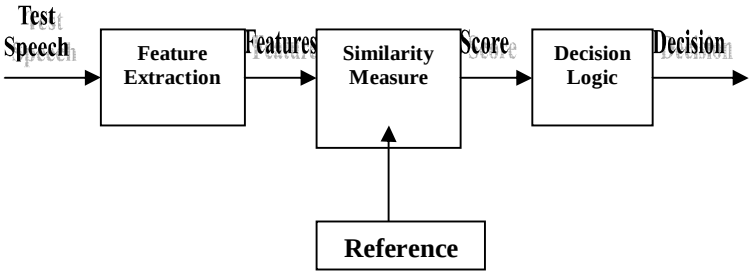

Fig : Testing phase [7]

3) Modeling Techniques : It's main objective s to generate speaker models using speaker specific feature vector. It is categorized into : speaker identification (which automatically identifies who is speaking on basis of information in speech signal) and speaker recognition. Speaker recognition is further classified as speaker dependent (which extract speaker characteristics to identify speaker) and speaker independent (which focuses on the content of the message rather than individual speaker).

Following are the various modeling techniques :

i) Acoustic-Phonetic Approach - In this appropriate labels are assigned to speech sounds as it postulates that phonemes in spoken language are characterized based on acoustic properties.

ii) Pattern Recognition Approach - It makes use of a mathematical framework and a training algorithm to obtain speech patterns. This step is called pattern training. Then unknown speeches are compared with these patterns to determine their identity in pattern comparison stage.

iii) Template Based Approach - The main idea is that certain speech patterns (or pre-recorded words) are stored, that are considered to be dictionary of candidate words. Recognition is then carried out by matching utterances to be identified with these templates and selecting the pattern that matches the best. Although segmentation on basis of phonemes is avoided, but producing and storing templates per word is impractical.

iv) Dynamic Time Wrapping (DTR) - Similarities between two utterances that vary in speed are measured with DTR. Often used in HMM and template based approach matching, it stresses and compresses various sections of speech so as to find the best possible match between template and utterances on frame-by-frame basis. The utterances are "wrapped" non-linearly in time dimension to determine their measure of similarity. Continuity factor plays a major role in DTW approach.

v) Knowledge Based Approach – Expert knowledge about variations in speech is hand coded in the system. However, this expert knowledge is hard to obtain and use effectively.

vi) Statistical Based Approach - Using statistical learning procedures, variations in speech are modeled statistically. This is the scenario in HMM and it has been quite successful too. K-means algorithm is also used where training vectors are divided into clusters to get K feature vectors at the end.

vii) Learning Based Approach - In this, the system automatically learns domain expert knowledge through emulations or evolutionary process.

viii) Artificial Intelligence Approach - It is a hybrid of acoustic phonetic and pattern recognition approach. Acoustics phonetic knowledge is used to develop speech classification rules. However, just like knowledge based approach, quantifying expert knowledge is difficult.

ix) Stochastic Approach - This approach is useful to deal with the uncertainties of speech recognition like confusable sounds, speaker verifiability and homophone sounds.

4) Matching Techniques :

i) Whole Word Matching – The incoming speech signal I compared to a pre-recorded template of words. Though comparatively it requires less processing but a large storage for every word to be pre-recorded is it's basic necessity, which is impractical.

ii) Sub-word Matching - Pattern recognition is based on phonemes. Though it requires comparatively more processing but having a much less storage requirement is it's plus point.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

III. HIDDEN MARKOV MODEL (HMM) : POPULAR PATTERN RECOGNITION MODELING TECHNIQUE

Out of the three approaches of speech recognition i.e. Acoustic-phonetic approach, Knowledge-based approach and Pattern recognition approach; HMM is a popular pattern recognition technique.

A Definition

HMM [6] is a model in which the sample system under consideration is a Markov process with hidden states. In other words, the state is not visible directly but output dependent on the state is visible. Mathematically, HMM is formulated as a triple: (A, B, π) where A is transition probability B is output emission probability π is initial state probability At time t, the model is in one state and at time t1, model moves to another state or stays in the same state and emits another observation. These transitions from one state to the other are only probabilistic. Transitions in speech recognition system can go from only left to right i.e. process cannot go backwards, thus enforcing temporal ordering of speech sounds.

HMM is characterized by :

1) N number of states in the model
2) M number of distinct observation symbols per state
3) S representation of individual state
4) Transition probability distribution A= {*aij*} where each aij is transition probability from S_i to state S_j

B Steps in HMM

1) Evaluation : It is the process of finding probability of generation of given observation sequence by a given model, which selects the best model from all the competing models.

$P(O \mid \lambda)$ means probability of observation sequence O given the HMM model λ. *Decoding* : The single best state sequence, Q, for given observation sequence is decoding, where $Q = (q_1, q_2, \ldots, q_T)$ and $O = (o_1, o_2, \ldots, o_T)$.

2) Training (Learning) : The most difficult step of HMM, it involves adjusting the model parameters (A, B, II) to maximize probability of observation sequence. Baum-Welch algorithm is used for this wherein we start with some initial estimates of the model parameters and modify them to maximize training observation sequence in iterative manner till a critical value is reached.

IV. MEL FREQUENCY CEPSTRUM COEFFICIENT (MFCC)

MFCC [7] is a popular text dependent speaker identification technique, in other words, it is used for feature extraction.

A. The technique - MFCC is a technique to extract and select the best parametric representation of acoustic signal for speech recognition. MFCC calculation involves the following steps:

1) Mel-frequency Wrapping : A subjective pitch is measured on 'mel' scale as humans do not follow a linear scale for frequency contents of sounds for frequency content of sounds for speech signal. Mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, pitch of 1 KHz tone, 40 dB above perceptual hearing threshold, is defined as 1000 mels. Mathematically, Mel(f) = 2595*log10(1+f/700)

2) Cepstrum : The log mel spectrum, being real numbers, are converted back to time which provides a good representation of local spectral properties of the signal; discrete cosine transform is done for the transformation :

$Cn = \sum (\log S_k) \cos \{n(k-(1/2)* \pi/k)\}$ *k=1*

V. DIFFICULTES FACED BY SPEECH RECOGNITION AND PROBABLE SOLUTIONS

A Difficulties

1) The way we pronounce words is affected by the words before and after it, this is called co-articulation. This also affects sounds within words.
2) Background noise also affects speech.
3) Extraneous noise may be passed to the speech recognizer.
4) Sound of the user might not be accurately conveyed to the recognizer.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

B Ways to Minimize Recognizer Errors

1) Constraint the recognizer's vocabulary i.e. limit the dictation speaker feeds to the recognizer.
2) "Noise Canceling" microphones can be used to limit background noise.
3) The recognizer can gain user's voice type information by letting him/her to speak to the dictation system for five to fifteen minutes prior to using the system.
4) Recognizer's can also speak "anchor words" pre, or pr and post each utterance. Though some aspects of performance will be improved but productivity is advertised affected.

VI. PERFORMANCE OF SYSTEMS

Performance of speech recognition system is measured in terms of accuracy (with word error rate) and speed (with real time factor). [7]

A Word Error Rate (WER)

$$WER = \frac{S+D+I}{N}$$

where S is number of substitutions D is number of deletions I is number of insertions N is number of words in the reference

VII. CONCLUSION

Advantages [4]

1) Since no special hardware is required so cost of implementation is low; a telephone or a microphone is all that is needed for authentication.
2) It is easy to use as speaking is natural.
3) It allows users to authenticate remotely.
4) Authentication by comparing the voiceprint created at enrollment and sample given by user is very fast, 0.5 seconds only is all that it needs.
5) The storage size needed by speech pattern is very small.

B Disadvantages [4]

Speech is highly influenced by factors such as :

1) Background noise 2) Age 3) Weather 4) Physical obstructions like cold, cough .

REFERENCES

[1] Pradeep Kumar Jaisal, Pankaj Kumar Mishra, "A Review of Speech Pattern Recognition Survey", International Journal of Computer Science and Technology, 2012.
[2] Santosh K.Gaikward, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications, 2010.
[3] "Speech Recognition Technology Choices", A Vocollect White Paper, 2010.
[4] Lisa Myers, "An Exploration of Voice Biometrics", SANS Institute Infosec, 2004.
[5] Rudan Bettelheim, David Steele, "Speech and Command Recognition", FreeScale White Paper, 2010.
[6] Bhupinder Singh, Neha Kapur, Puneet Kaur, "Speech Recognition with Hidden Markov Model : A Review", International Journal of Advanced Research in Computer Science and Software Engineering, 2012.
[7] Vibha Tiwari, "MFCC and it's Applications in Speaker Recognition", International Journal on Emerging Technique, 2010.
[8] Vimala C., Dr. V.Radha, "A Review on Speech Recognition Challenges and Approaches", World of Computer Science and Information Technology Journal, 2012. A.P.
[9] Henry Charles G. Devaraj, "Alaigal- A Tamil Speech Recognition", Tamil Internet Singapore, 2004.
[10] Er. Jaspreet Kaur, Er. Nidhi, Ms. Rupinder Kaur, "Issues Involved in speech To Text Conversion", International Journal of Computional Engineering Research, 2012.
[11] Kuldeep Kumar, R.K.Aggarwal, "Hindi Speech Recognition System using HTK", International Journal of Computing and Business Research, 2011.

BIOGRAPHY

Anjum Asma Mohammed is a Research Assistant in the Information Technology Department, College of Computer and Information Sciences, King Saud University. She received Master of Computer Application (MCA) degree in 2005 from BAMU, Aurangabad, MS, India. Her research interests are Computer Networks (wireless Networks), HCI, Algorithms, web 2.0 etc.