International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347- 5552, Volume-2, Issue-1, January-2014

Review of Speech Recognition System

Jyoti Madan, Ajmer Singh

***Abstract*- This paper takes a tour of speech recognition system which includes it's evaluation and accuracy of system and discuss the structure of utterance that uses the vocal tract to make the utterance. Dynamic warping with its neural approach to convert the speech into text. This paper also explains the basic working of speech recognition system with elaboration of it's techniques.**

***Keywords*- Speech Recognition System, Dynamic Time Warping, Neural Network, Automatic Speech Recognition(ASR) Performance.**

I INTRODUCTION

Speech is a natural mode of communication for people. We learn all the relevant skills during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. It comes so naturally to us that we don't realize how complex a phenomenon speech is. The human vocal tract and articulators are biological organs within onlinear properties, whose operation are not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. As a result, vocalizations can vary widely in terms of their accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; moreover, during transmission, our irregular speech patterns can be further distorted by background noise and echoes, as well as electrical characteristics.

A. Types of Speech [1][5][6]

Speech recognition systems differ on the basis they accept utterances as input.

***1. Isolated Words*:** These recognizers require input to be in the form of isolated utterances, i.e. the speaker must wait in-between utterances in order to give recognizer time to do the required processing.

2. Connected Words: Connected utterances are fed as input in these recognizers, i.e. the speaker is required to speak separate utterances together with minimum pause in-between them.

Manuscript received January 5, 2014
Jyoti Madan, Research candidate, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, (jyoti0038@gmail.com)
Ajmer Singh, Asst. Professor, Department of Computer Science and Engineering, , Deenbandhu Chhotu Ram University of Science and Technology, Murthal, (ajmer.saini @gmail.com)

3. Continuous Speech: A computer dictation to the user, this type of recognizer is considered to be the most difficult to create.

4. Spontaneous Speech: This recognizer takes the natural speech of speaker as input.

B. Evaluation and accuracy of the system:

- **Vocabulary size and confusability:-** it is easy to discriminate among a small set of words, but error rates naturally increase as the vocabulary size grows.
- **Speaker Dependence vs. Independence:-**A speaker dependent system is intended for use by a single speaker, but a speaker independent system is intended for use by any speaker. Speaker independence is difficult to achieve because a system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific.
- **Isolated, discontinuous, or continuous speech: -** Isolated speech means single words; discontinuous speech means full sentences in which words are artificially separated by silence; and continuous speech means naturally spoken sentences. Isolated and discontinuous speech recognition is relatively easy because word boundaries are detectable and the words tend to be cleanly pronounced.
- **Task and language constraints** :- Even with a fixed vocabulary, performance will vary with the nature of constraints on the word sequences that are allowed during recognition. Some constraints may be task-dependent.
- **Read vs. spontaneous speech:-** Systems can be evaluated on speech that is either read from prepared scripts, or speech that is uttered spontaneously. Spontaneous speech is vastly more difficult.

II. SPEECH RECOGNITION SYSTEM

Speech recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of subword units, words, phrases, and sentences.

Each level may provide additional temporal constraints.
This hierarchy of constraints can best be exploited by combining decisions probabilistically at all lower levels, and making discrete decisions only at the highest level.

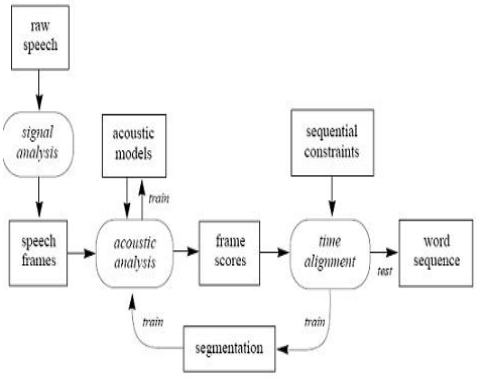


Fig 1: Structure of Speech Recognition System

- **Raw speech**:- Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time.
- **Signal analysis**:- Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information.
  - Fourier analysis (FFT) yields discrete frequencies over time, which canbe interpreted visually. Frequencies are often distributed using a Mel scale, which is linear in the low range but logarithmic in the high range,corresponding to physiological characteristics of the human ear.
  - Perceptual Linear Prediction (PLP) is also physiologically motivated, but yields coefficients that cannot be interpreted visually.
  - Linear Predictive Coding (LPC) yields coefficients of a linear equation that approximate the recent history of the raw speech values.
  - Cepstral analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the signal.

In practice, it makes little difference which technique is used1. Afterwards, procedures, such as Linear Discriminant Analysis (LDA) may optionally be applied to further reduce the dimensionality of any representation, and to decorrelate the coefficients.

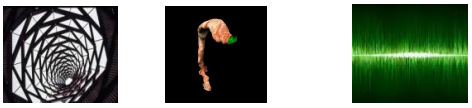
A. Dynamic Time Warping

Dynamic Time Warping algorithm is one of the oldest and most important algorithms in speech recognition. The simplest way to recognize an isolated word sample is to compare it against a number of stored word templates and determine which is the “best match”. This goal is complicated by a number of factors. First, different samples of a given word will have somewhat different durations. This problem can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration. However, another problem is that the rate of speech may not be constant throughout the word; in other words, the optimal alignment between a template and the speech sample may be nonlinear. Dynamic Time Warping (DTW) is an efficient method for finding this optimal nonlinear alignment.
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), computer speech recognition", "speech to text", or just "STT". Speech recognition systems are proving to be beneficial as not only human can speak more quickly instead of typing on a keyboard but also speech input provides relative ease of use. Speech recognition systems vary on basis of the way they accept utterances as input, these ways are: isolated words, connected words, continuous speech or spontaneous speech.

B. Structure and Experimental Study of Utterance [2]

An utterance is converted into its structure, which is a holistic and speaker-invariant representation of the utterance. If we want to convert the structure back to sounds again, we have to specify the size of the vocal tract based on who will realize this structure in an actual sound (voice) space. aims at realizing a

structure in an actual sound space without that interface. However, it is true that, only with a structure, this process is impossible to execute because a structure contains no information at all on where in the sound space each event (distribution) of a BD-based matrix should be realized. Here, we introduce acoustic instances of a few events of the matrix as initial conditions.



Structure + Vocal Tract = Utterance

Fig 2: Structure of Utterance [2]

The invariance experimentally and showed the super robustness of the proposed framework for ASR. The task was recognizing isolated words and the words were defined as sequences of 5 Japanese vowels such as /aeoui/. It is well-known that vowel sounds are much more dependent on speakers than consonant sounds. Since Japanese has only 5 vowels, the vocabulary size was 120. Utterances of 4 male and 4 female adult speakers were used to train two recognizers, one was with word-HMMs and the other was with structures.

C. Neural Network Approach [15]

Neural networks are good at pattern recognition, many early researchers naturally tried applying neural networks to speech recognition. The earliest attempts involved highly simplified tasks, e.g., classifying speech segments as voiced/unvoiced, or nasal/fricative/plosive.

There are two basic approaches to speech classification using neural networks: static and dynamic, as illustrated in Figure.

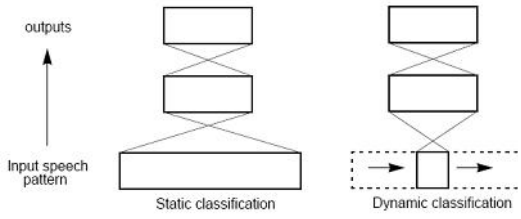


Fig 3: Static and dynamic approaches to classification

In static classification, the neural network sees all of the input speech at once, and makes a single decision. By contrast, in dynamic classification, the neural network sees only a small window of the speech, and this window slides over the input speech while the network makes a series of local decisions, which have to be integrated into a global decision at a later time. Static classification works well for phoneme recognition, but it scales poorly to the level of words or sentences; dynamic classification scales better. Either approach may make use of recurrent connections, although recurrence is more often found in the dynamic approach.

D. ASR Performance [1][5][6]

On the one hand an ideal ASR performance index should reflect human judgement, which will depend in turn on the ASR application. On the other, an acceptable metric for ASR system evaluation must be simple to apply and not be language or application dependent. An intuitively appealing measure would be the proportion of (Shannon) information communicated, but this would depend on the model used for encoding, and natural language encoding makes strong use of word context, language specific grammatical structure and complex pragmatics. If we are to measure information, then we must therefore settle for a model of speech encoding which is context free, i.e. in which all that is being communicated is a series of independent words.

The two basic parameters on basis of which the speech recognizer system is evaluated are: WER (Word Error Rate) and RTF (Real Time Factor). Word error rate (WER) is a metric to measure the performance of a speech recognition in terms of it's accuracy. WER is used for comparing different systems as well as for evaluating improvements within one system. Real time factor (RTF) is used to measure the speed of an ASR system i.e. the rate at which conversion takes place. WER is given as :

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

where S is number of substitutions, D is number of deletions, I is number of insertions and N is number of words in the reference. The current speech recognizer systems are such

affected by this factor that if we handle the way insertions are made then deletions will cause a problem for the correct speech recognition. So overall a balance needs to be maintained between all of these factors to obtain the correct or rather accurate result.

RTF is given as :

$$RTF = \frac{P}{I}$$

where P is time taken to process an input of duration I. This is a parameter that needs special attention. For any system to be considered as effective, it must process the input and produce the output within as far as possible a short time interval. So grammar rules made should be such that the appropriate match is found and that too very quickly.

III. BASIC WORKING [1]

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. 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 w**ords**

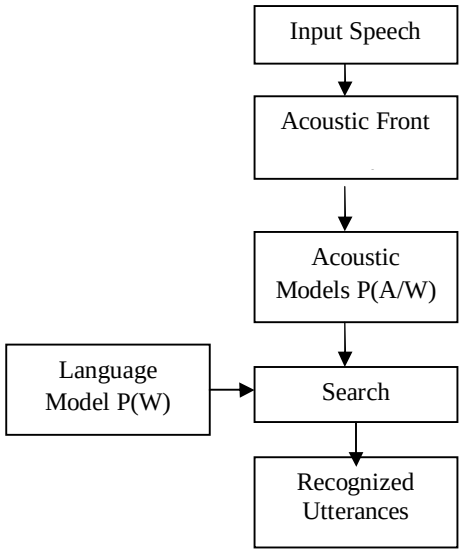


Fig 4: Working Principle of Speech Recognition System

A. Technique [6]

i) **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.

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

iii) Modeling Techniques :
It's main objective is 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).

iv) Matching Techniques :
i) Whole Word Matching – The incoming speech signal is 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.

IV. CONCLUSION

Research in speech recognition system has been in existence for more than 50 years as not only these applications are handy for the persons with disabilities, who are unable to make use of keyboard or mouse to operate computer, but also these ASR systems can be developed in one's native language like English, Punjabi, Tamil, Hindi, Chinese, etc; which ultimately breaks the barrier of the person being educationally underprivileged to operate the computer with explanation of oldest algorithm of Dynamic Time Warping. There is an experimental study of utterance and provide a structure of it with vocal tract. Large vocabulary, speaker independent continuous speech recognition systems are in-demand which can be fulfilled. Speech Recognition is related with Neural Network approach that helps to convert the speech into text with continuous speech.

V. FUTURE SCOPE

The future systems must be trained in a manner that they can withstand a robust environment so that their performance doesn't degrade i.e. be speaker independent. Moreover, the systems must be trained for new task so that they do not suffer from degradation, irrespective of time and money consumed i.e. trained to adapt themselves as per any language. Systems must also be capable enough to deal with spontaneous speech phenomena such as filled pauses, ungrammatical constructions, false starts and hesitations.

REFERENCES

[1] Andrew C. Morris, Viktoria Maier & Phil Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition" , Institute of Phonetics Saarland University, Germany, 1999.

[2] Seongjae Lee, Sunmee Kang, Hanseok KO, Jongseong Yoon, Minseok Keum, "Dialogue Enabling Speech–to– Text User Assistive Agent with Auditory Perceptual Beam forming for Hearing-Impaired" , International Conference on Consumer Electronics (ICCE) IEEE, 2013.

[3] Nobuaki Minematsu, Daisuke Saito, Keikichi Hirose, "Experimental Study of Structure to speech Conversion" , ICSP Proceedings IEEE, 2008.

[4] Gyo¨rgy Szasza´k, A´ kos Ma´te´ Tu¨ndik and Kla´ra Vicsi, "Automatic Speech to Text Transformation of Spontaneous Job Interviews on the HuComTech Database", Budapest University of Technology and Economics, 1997.

[5] Pradeep Kumar Jaiswal, Pankaj Kumar Mishra, "A Review of Speech Pattern Recognition Survey" International Journal of Computer Science and technology, 2012, ISSN 0976-8491 (Online) | ISSN : 2229-4333.

[6] Abhay Bansal, Kamal kant, Komal Chauhan, "A Review of Speech Recognition System", COMPTECH: An International Journal of Computer Sciences(ISSN: 2229-4589)(Global Publication: Delhi) Vol. 3, No. 6(January 2013).

[7] Alex Acero, "An Overview of Text-to-Speech Synthesis", Speech Technology Group Microsoft Corporation Redmond, WA 98052, 0-7803-64 16-3/00/2000 IEEE.

[8] Lori Lamel, Jean-Luc Gauvain, Viet Bac Le, Ilya Oparin, Sha Meng, " Improved Models For Mandarin

Speech-To-Text Transcription", Spoken Language Processing Group, LIMSI-CNRS 91403 Orsay, FRANCE, 978-1-4577-0539-7/11/2011 IEEE.

[9] Jordi Adell, Pablo Daniel Ag¨uero, Antonio Bonafonte,"Database Pruning For Unsupervised Building Of Text - To-Speech Voices", Dept. of Signal Theory and Communications , TALP Research Center, 142440469X/06/2006 IEEE.

[10] Christine Tuerk, Peter Monaco and Tony Robinson, "The Development Of A Connectionist Multiple-Voice Text-To-Speech System", Cambridge University Engineering Department, CH2977-719 11000-0749@ 1991 IEEE.

[11] Amarasekara M.S., Bandara K.M.N.S., Yithana B.Y.A.I., 0e Silva O.H., Jayakody, " Real Time Interactive Voice Communication ", The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka, 978-1-4673-4463-0/13/$31.00 ©2013 IEEE.

[12] Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, Herve Bourlard, "On The Use Of Information Retrieval Measures For Speech Recognition Evaluation", March 2005, IDIAP{RR 04-73.

[13] Ralph Alter, "Utilization Of Contextual Constraints In Automatic Speech Recognition", TRSNSACTIONS ON AUDIO AXD ELECTROACOUSTICS VOL. AU-16, NO. 1 MARCH 1968, IEEE.

[14] F. Diehl, M.J.F. Gales, M. Tomalin, & P.C. Woodland,"Phonetic Pronunciations For Arabic Speech-To-Text Systems", Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K., 1-4244-1484-9/08/ ICASSP 2 008 IEEE.

[15] L.Srivathsan, R.Srikanth, S.Sivasankaran, M.Chandru, "NEURAL SPEECH - An Aid For Differently Challenged", Instrumentation & Control Department, Sri Sairam Engineering College, Anna University, International Conference on Computational Intelligence and Computing Research,2012, IEEE.