Automatic Speech Recognition: A Review

Anchal Katyal, Amanpreet Kaur, Jasmeen Gill

Abstract - The research, development and the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges over the years e.g. speaker and language variability, vocabulary size and domain, noise. This paper describes the recent progress and the author's perspective of ASR and gives an overview of major technological perspective and appreciation of the fundamental progress of Automatic speech recognition.

Index Terms - ASR, HMM, Classifications of ASR, Speech Recognition Process, HNN.

I. INTRODUCTION

Speech recognition is a computer science term and is also known as automatic speech recognition. It is a feature that turns speech into text. One of the main advantages for speech recognition services is the cut back on misspelled words that some typists may suffer from when typing. The service cuts down on the amount of time editing and fixing spelling corrections [1] [3]. It is also a big advantage to people who may suffer from disabilities that affect their writing ability but can use their speech to create text on computers or other devices. The overall advantage is the time management. Most people can speak faster than they can type with fewer mistakes.

1.1 Types of speech recognition

A. Text-To-Speech

Text-To-Speech (or TTS) will manipulate a string of text into an audio clip. It is useful for blind people to be able to use computers but can also be used to simply improve computer experience. There are several programs available that perform TTS, some of which are command-line based (ideal for scripting) and others which provide a handy GUI [2].

B. Simple Voice Control/Commands

This is the most basic form of Speech-To-Text application. These are designed to recognize a small number of specific, typically one-word commands and then perform an action. This is often used as an alternative to an application launcher, allowing the user for instance to say the word *"Firefox*" and have his OS open a new browser window [1].

C. Full dictation/recognition

Full dictation/recognition software allows the user to read full sentences or paragraphs and translates that data into text on the fly. This could be used, for instance, to dictate an entire letter into the window of an email client [3]. In some cases, these types of applications need to be trained to your voice and can improve in accuracy the more they are used.

II. AUTOMATIC SPEECH RECOGNITION (ASR)

It means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation [8] [4].In recent years, automatic speech recognition technology has advanced to the point where it is used by millions of individuals to automatically create documents from dictation. One of the major problems in automatic speech recognition technologies is the sensitivity of recognizers to any interfering sounds. Since natural environments often include other sound sources, the performance of the existing technologies is severely limited. For a phonetic language, there always exits a one to one mapping between their pronunciation and orthography [2]. In addition to this, unlike English and other European languages, these languages possess a large number of phonemes such as retroflex and aspirated stops. Some of the major application areas of Automatic speech recognition systems are dictation, controlling the programs, automatic telephone call processing and query based information system such as travel information system , weather report information system etc. In simple words, Speech Recognition is the process to take the audio format as input and then convert the text format from it as output [8].

2.1 Benefits of ASR

- Accessibility for the deaf and hard of hearing
- Cost reduction through automation
- Searchable text capability

2.2 Mathematical representation of ASR

In statistical based ASR systems an utterance is represented by some sequence of acoustic feature observations O, derived from the sequence of words W. The recognition system needs to find the most likely word sequence, and given the observed acoustic signal is formulated by [1]:

$$W = \mathrm{argmaxW}\ P(W|O) \qquad (i)$$

In "equation (i)" the argument, P(W|O) i.e. the word sequence W is found which shows maximum probability, given the observation vector O[1][2]. Using Baye's rule it can be written as:

$$W = \mathrm{argmaxW}\ P\ (W|O).\ P\ (W)/P\ (O) \qquad (ii)$$

In "equation (ii)",P (O) is the probability of observation sequence and is not considered as it is a constant w.r.t. W. Hence,

$$W = argmaxW\ P\ (W|O)\ P(W) \qquad (iii)$$

In "equation (iii)", P (W) is determined by a language model like grammar based model and P(O|W) is the observation likelihood and is evaluated based on an acoustic model.

Manuscript received February, 2014.

Anchal Katyal, M.Tech Scholar (CSE) RIMT, Mandi Gobindgarh
Ms.Amanpreet Kaur, AP (CSE) BBSBEC, Fatehgarh Sahib
Ms. Jasmeen Gill, AP (CSE) RIMT-IET, Mandi Gobindgarh

Among the various models, Hidden Markov Model (HMM) is so far the most widely used technique due to its efficient algorithm for training and recognition.

2.3 Phases of ASR

Automatic speech recognition system involves two phases:

A. Training phase

A rigorous training procedure is followed to map the basic speech unit such as phone, syllable to the acoustic observation. In training phase, known speech is recorded, pre-processed and then enters the first stage i.e. Feature extraction. The next three stages are HMM creation, HMM training and HMM storage [2].

B. Recognition phase

The recognition phase starts with the acoustic analysis of unknown speech signal. The signal captured is converted to a series of acoustic feature vectors. Using suitable algorithm, the input observations are processed. The speech is compared against the HMM's networks and the word which is pronounced is displayed. An ASR system can only recognize what it has learned during the training process. But, the system is able to recognize even those words, which are not present in the training corpus and for which sub-word units of the new word are known to the system and the new word exists in the system dictionary [4].

III. CLASSIFICATIONS OF ASR

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech [7].

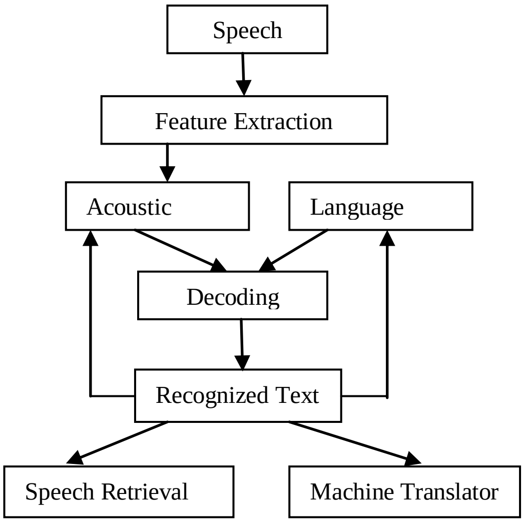

Fig1: Main Components of an Automatic Speech Recognition System

A. Feature Extraction

Signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other.

B. Acoustic modeling

It provides probabilities for different phonemes at different time instants. It is the statistical mapping from the units of speech to all the features of speech. These are used for speech sounds to phoneme and from phoneme to word.

C. Language modeling

It defines what kind of phoneme and word sequences are possible in the target language or application at hand and what are their probabilities.

D. Decoding

The acoustic models and language models are used in for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

IV.SPEECH RECOGNITION PROCESS

In essence, the basic task involved in speech recognition is that of going from speech recordings to word labels. As the pattern recognition approach to speech recognition is the most widely used approach. There are two main variants of the basic speech recognition task, namely isolated word recognition and connected word recognition [2] [3].

4.1 Variants of the Speech Recognition Task

A. Isolated word recognition

Isolated word recognition refers to the task of recognizing a single spoken word where the choice of words is not constrained to task syntax or semantics. HMMs can be used to build an isolated word recognizer. HMM approach is a well-known and widely used statistical method of characterizing the spectral properties of the frames of a pattern. HMMs are particularly suitable for speech recognition as the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be determined in a precise, well-defined manner.

B. Fluent speech Recognition

Fluent speech recognition is a more complicated task than isolated word recognition. In this case the task is to recognize a continuous string of words from the vocabulary.

C. Feature Extraction and Pattern Recognition

The input into an automatic speech recognition system is the speech signal. The two major tasks involved in speech recognition are feature extraction and pattern recognition.

Feature Extraction

In all speech recognition systems the first step in the process is signal processing. Initially a spectral / temporal analysis of the speech signal is performed to give observation vectors which can be used to train the HMMs. One way to obtain observation vectors from speech samples is to perform spectral analysis. A type of spectral analysis that is often used is linear predictive coding.

Pattern Recognition

Pattern recognition refers to the matching of features. The pattern recognition process consists of training and testing. During training, a model of each vocabulary word must be created. Each model consists of a set of features extracted from the speech signal. The exact form of the model depends on the type of pattern-recognition algorithm used. During testing, a similar model is created for the unknown word [3]. The pattern-recognition algorithm compares the model of the unknown word with the models of known words and selects the word whose model score is highest. There are many different pattern matching techniques.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

These include templates, Dynamic Time Warping and HMMs.

V. HIDDEN NEURAL NETWORKS

A general framework for hybrids of hidden Markov models (HMMs) and neural networks (NNs) called hidden neural networks (HNNs) [16]. A neural network model consists of units with associated basis or activation functions, which are connected through weights. These units are arranged in layers. Although ANN structures like the radial basis function networks exist where the parameters of the activation functions are estimated during training, the ANNs used in this thesis use fixed activation functions. The connecting weights are the free parameters to be estimated. Every ANN has an input layer with one unit for each dimension of the observation vector and an output layer with one unit for each target class. Within these layers one or more layers are included, which have no particular desired behavior and have no direct connection to the outer world. For this reason they are called hidden layers. Hidden units are able to discover regularities in the data and enrich the family of functions the network is able to approximate [7].

A. Hidden Markov Model

Hidden Markov models are special cases of neural networks [16]. Despite huge amounts of research trying to create an intelligent speech recognition machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments. To date, the best results in speech recognition systems have been achieved by those based on hidden Markov models. Hence, most current automatic speech recognition systems are based on HMMs.

B. Three Fundamental Problems of HMM design

HMM design is characterized by three fundamental problems namely:

- The evaluation of the probability of a sequence of observations given a specific HMM.
- The determination of a best sequence of model states.
- The adjustment of model parameters to best account for the observed signal.

There are various methods for solving the above problems. The most popular technique used to solve problem 1.

The Forward-Backward procedure, is an algorithm for computing the probability of a particular observation sequence.

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states that results in a sequence of observed events. This algorithm is a popular technique for solving problem 2, that of finding the best state sequence for the given observation sequence [15].

The third and most difficult problem in the design of HMMs is the problem of determining a method to maximise the probability of the observation sequence given the model. As Rabiner mentions, there is no known way to analytically solve this problem, neither is there an optimal way of estimating the model parameters. There are various iterative procedures such as the Baum-Welch method, and expectation modification method or gradient techniques that can be used to choose model parameters. The standard criterion for estimation of HMM parameters is maximum likelihood.

C. Types of HMMs

There are many different types of hidden Markov models. In the ergodic or fully connected HMM, every state of the model can be reached (in a single step) from every other state of the model. Inspeech processing, the left-right model or Bakis model has been used. The benefit of this model is that it can model signals whose properties change over time [7] [16].

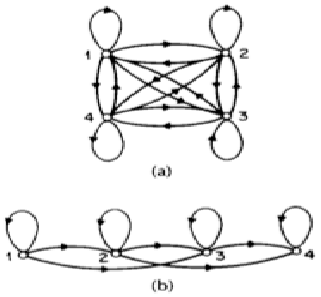

Fig.2: Illustration of 2 types of HMMs. (a) A 4-state ergodic model. (b) A 4-state left-right model

VI.CONCLUSION

Automatic Speech Recognition is the challenging problem to deal with. In this paper, we discussed the various techniques of Automatic Speech Recognition and Hidden Markov Model (HMM) of how the technology has progressed from the last years. ASR is more than automatic text to speech; ASR requires fast computers with lots of data capacity and memory a necessary condition for complex recognition tasks, and the involvement of speech scientists, linguists, computer scientists, mathematicians, and engineers.

REFERENCES

[1]. Giuseppe Riccardi, Dilek Hakkani-tur, “ Active learning: theory and applications to Automatic speech recognition,” IEEE transactions on speech and audio processing, vol. 13, no. 4, july 2005.

[2]. Santosh K.Gaikwad, Bharti W.Gawali, “A review on speech recognition technique,” International Journal of computer applications (0975 – 8887) Volume 10– no.3, November 2010.

[3]. Thomas Hain, Asmaa El Hannani, “Automatic Speech Recognition for scientific purposes – webasr,”2008 ISCA, september 22-26, brisbane australia.

[4]. Parwinderpal Singh, Er. Bhupinder Singh, “ Speech recognition as emerging revolutionary technology,”International Journal of advanced research in computer science and software engineering Volume 2, issue 10, october 2012 issn: 2277 128x.

[5]. Wiqas Ghai, Navdeep Singh, “Analysis of automatic speech recognition systems for indo-aryan languages: Punjabi a case study,” International Journal of Soft computing and engineering (IJSCE) issn: 2231-2307, volume-2, issue-1, march 2012.

[6]. Preeti Saini, Parneet Kaur, “Automatic Speech Recognition: a review,” International Journal Of Engineering Trends And Technology- Volume4issue2- 2013.

[7]. Mohit Dua, R.K.Aggarwal, “ Punjabi Automatic Speech Recognition using htk,” IJCSI international journal of computer science issues, vol. 9, issue 4, no 1, july 2012 issn (online): 1694-0814.

[8]. Wiqas Ghai, Navdeep Singh, “Literature review on automatic speech recognition,” international journal of computer applications (0975 – 8887) volume 41– no.8, march 2012.

[9]. I. A. Muslea, “active learning with multipleviews,” ph.d. Dissertation, Univ. Southern california, los angeles, 2000.

[10]. M. Tang, x. Luo, and s. Roukos, “active learning for statistical natural Language parsing,” in proc. 40th anniversary meeting of association For computational linguistics (acl-02), philadelphia, pa, 2002, Pp. 120–127.

[11]. G. Riccardi and d. Hakkani-tür, “active and unsupervised learning for Automatic speech recognition,” in proc. Eurospeech, 2003.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[12]. M.a.anusuya , s.k.katti “speech recognition by machine :A review” international journal of computer science and Information security 2009.

[13]. Samudravijay k “speech and speaker recognition report” Source:http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html Viewed on 23 feb. 2010.

[14]. Sadokifuruki ,Tomohisa ichiba, “Cluster-based modeling For ubiquitous speech recognition, Tokyo institute of technology Interspeech 2005.

[15] Spector, Simon kinga and Joe Frankel, “Recognition ,speech Production knowledge in automatic speech recognition ,Journal of acoustic society of america,2006.

[16]. Ranu Dixit,Navdeep Kaur, “Speech Recognition Using Stochastic Approach: A Review,” International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 2, February 2013.

Author 1: Jasmeen Gill, a dynamic lecturer. She has completed M. Tech and B Tech in Computer Science Engineering having an experience of 8 years. She has published 3 national and international papers. Her areas of interest are Artificial Intelligence and Image Processing. She has a membership of IEEE.

Author 2: AmanPreet Kaur, an effectual lecturer has an experience of 10.5 years. She has completed M. Tech and B Tech in Computer Science Engineering and pursuing P.hd. She has published 10 national and international papers and pursuing research work on Punjabi Speech recognition. She has a membership of IEEE.

Author 3: Anchal Katyal has completed B.Tech and Pursuing M.Tech.She has published four national and international papers.She has a membership of IEEE.