International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Review on Automatic Speech Recognition

C. Venish Raja*[1], M. Daisy Jackuline[2], M. Ranjitha[3]

*[1]Assistant Professor, Department of Information Technology, St.Joseph's College, Trichy, Tamilnadu, India
2M.Sc CS, Department of Information Technology, St.Joseph's College, Trichy, Tamilnadu, India
[3]M.Sc CS, Department of Information Technology, St.Joseph's College, Trichy, Tamilnadu, India

ABSTRACT

With the advancement of speech recognition technologies, there is an increase in the adoption of voice interfaces on mobile-based platforms. While, developing a general purpose Automatic Speech Recognition (ASR) which can understand voice commands is important, the contexts of how people interact with their mobile device change very rapidly. Due to the high processing complexity of the ASR engine, much of the processing of trending data is being carried out on cloud platforms. Changed content regarding news, music, movies and TV series change the focus of interaction with voice based interfaces. Hence ASR engines trained on a static vocabulary may not be able to adapt to the changing contexts. The focus of this paper is to first describe the problems faced in incorporating dynamically changing vocabulary and contexts into an ASR engine. We then propose a novel solution which shows a relative improvement of 38 percent utterance accuracy on newly added content without compromising on the overall accuracy and stability of the system.

Keywords : Speech, Speech Recognition, Human Machine Interaction, Communication

I. INTRODUCTION

ASR is the first stage in an overall human/computer interaction pipeline that also includes Voice box's related Natural Language Understanding (NLU) and Text-to-Speech (TTS) [10] technologies. Voice box's advanced SR module is a multi-stage pipeline that uses techniques from machine learning, graph theory, traditional grammar development, and statistical analysis of large corpuses to form high-confidence transcriptions of input audio.

Real-world concerns everything from accent and dialect differences, to CPU and memory limitations on different platforms, to operation despite environmental noise, and more present additional challenges beyond what academic ASR research can achieve in controlled laboratory environments. Voice box's ASR system uses a wide range of techniques to overcome those obstacles and maintain robust performance under real-world conditions. Voice box offers ASR solutions for a wide range of CPU, memory, and connectivity limitations. The Voice box SR can support fully-embedded operation on small-footprint devices, hybrid combinations of embedded and server-based SR, and fully connected scenarios in which the SR runs on a high-performance server in the cloud.

Automated speech recognition (ASR) is a technology that allows users of information systems to speak entries rather than punching numbers on a keypad. Today, advances in mobile computing have put the power of ASR into billions of cell phones, automotive consoles, and Internet of Things[3][4] (IoT) devices world-wide.

The limitations of these platforms—notably, their lack of physical keyboards—has created a vast need for an easy, accurate, natural form of interaction with those devices. Using voice and ordinary language has been the dream of human/computer interaction since the idea was first introduced to a mass audience through the 1966 television series Star Trek.

Real advances in computer technology have far outstripped even the vision of Star Trek in terms of miniaturization, communications, and the overall ubiquity of computation. Yet, due to the practical challenges of speech processing, the dream of voice interaction has remained of science fiction for over half a century. But today, as the bridge between the user and Voice box's related technologies of Natural Language Understanding (NLU) and Text-to-Speech (TTS), ASR brings that dream to reality.

II. BASIC PRICIPLE OF VOICE RECOGNITION

The speech recognition system is essentially a pattern recognition system, including feature extraction, pattern matching and the reference model library [1] [10].

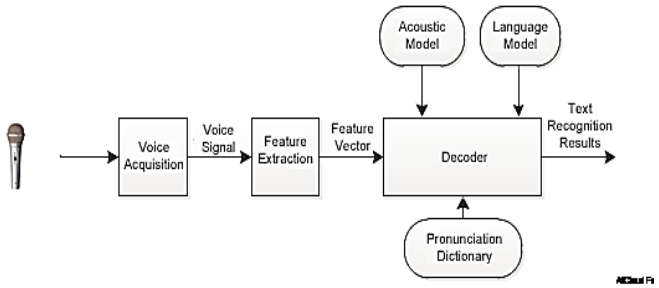

Figure 1: Voice Recognition Processing

The unknown voice through the microphone is transformed into an electrical signal on the input of the identification system, the first after the pre-treatment.

The system establishes a voice model according to the human voice characteristics analyzes the input voice signal and extracts the required features on this basis; it establishes the required template of the speech recognition. The computer is used in the recognition process according to the model of the speech recognition to compare the voice template stored in the computer and the characteristics of the input voice signal. Search and matching strategies to identify the optimal range of the input voice matches the template. According to the definition of this template through the lookup table can be given the recognition results of the computer [1]

2.1 Voice Acquition

The final way of reducing noise effects is to change the method of gathering the speech signal. In some environments head-set microphones can be used. These can be very directional and so pick up very little noise from outside sources, since they are in a fixed position with respect to the speaker. However, head- sets are not practical in all situations. They must be connected by a wire or radio link to the computer, they only work with one speaker, and they are fragile [3].

2.2 Feature Extraction

Feature extraction is the main part of the speech recognition system. It is considered as the heart of the system. The work of this is to extract those features from the input speech (signal) that help the system in identifying the speaker. Feature extraction compresses the magnitude of the input signal (vector) without causing any harm to the power of speech signal [12].

2.3 Acoustic Model

An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes upkoi78 a word. Each of these statistical representations is assigned a label called a phoneme. An acoustic model is created by taking a large database of speech and using special training

algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models ("HMM"s). Each phoneme has its own HMM [15] [16].

2.4 Language Model

Language models are used to constrain search in a decoder by limiting the number of possible words that need to be considered at any one point in the search. The consequence is faster execution and higher accuracy. A **Statistical Language Model** is a file used by a Speech Recognition Engine to recognize speech [14]. It contains a large list of words and their probability of occurrence. It is used in dictation applications.

2.5 Pronunciation Dictionary

In constructing a speech recognition system, pronunciation information must be provided for all words spoken in both test and training data. The pronunciation dictionary on first release contains over 96,000 word definitions.

III. TYPES OF SPEECH RECOGNITION

There are commonly three different types of recognition in use in different applications:

3.1 Grammer_driven
3.2 Dictation
3.3 Natural Language Understanding (NLU)

In improving the overall performance of the system by migrated the load dynamically.

3.1 Grammer_driven

Grammar-driven speech recognition works best when your application is looking for a specific piece of data, like a name, phone number, account number, dollar amount, or date. Think of this type of speech recognition like the small edit fields you find in a web page or mobile app for short answers. This is the only approach that will work reliably with a minimal amount of effort for names, technical terms, and non-common words that may be awkwardly spelled. The tradeoff for this accuracy is that your app needs to know ahead of time what a caller will say. The grammars in these apps can be complex definitions written in BNF or XML, but also in a simple list of phrases [11] [23].

3.2 Dictation

Dictation is used when you want a word-for-word transcription of what a user said. Dictation accuracy has improved significantly in recent years with the introduction of machine learning applied to massive repositories of recorded audio data, but recognition accuracy scores can vary greatly depending on the quality of the audio and the environment it was recorded in.

3.3 Natural Language Understanding (NLU)

NLU encompasses a range of technologies, but its most common uses are mapping the many ways humans speak to computer identifiable objects and intents, and sentiment analysis (is the user happy or angry?). An NLU application can ask broad questions and will try to determine the caller's intent whether they speak in complete sentences or sentence fragments. This can reduce the number of steps, or turns, that a user takes to reach their goal, and shorten call times.

IV. METHODS OF SPEECH RECOGNITION

4.1 Hidden Markov Model

Hidden Markov modeling is, as the name suggests, a modeling approach. Thus, there are three In addition the speech data can things to consider: the model, the method of computing the probability of the model giving rise to a particular output and the method of computing the parameters of the model from known examples of the word it is to represent.

A hidden Markov model (HMM) is a doubly stochastic process for producing a sequence of observed symbols [19]. An underlying stochastic finite state machine (FSM) drives a set of stochastic processes, which produce the symbols. When a state is entered after a state transition in the FSM, a symbol from that state's set of symbols is selected probabilistically for output. The term "hidden" is appropriate because the actual state of the FSM cannot be observed directly, only through the symbols emitted. In the case of isolated word recognition, each word in the vocabulary has a corresponding HMM. These HMMs might actually consist of HMMs that model sub word units such as phonemes connected to form a single word model HMM. In the case of continuous word recognition, a single HMM corresponds to the domain grammar. This grammar model is constructed from word-model HMMs. The observable symbols correspond to (quantized) speech frame measurements

4.2 Neural Network Model

The neural network model that more commonly used and has the potentiating of speech recognition mainly include single layer perception model, multi-layer perception model, Kohonen self-organizing feature map model, radial basis function neural network, predictive neural network, etc. In addition, in order to make the neural network reflects the dynamic of the speech signal time-varying characteristics, delay neural network, recurrent neural network and so on. A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs [5] [7].

4.3 Dynamic Time Warping

HMM corresponds to the domain grammar. This grammar model is constructed from word-model HMMs. The observable symbols correspond to (quantized) speech frame measurements [14].

A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g., time series) with certain restrictions. That is, the sequences are "warped" non-linearly to match each other. This sequence

Hidden models.

V. APPLICATIONS OF SPEECH RECOGNITION

5.1 Telecommunication Applications of Speech Recognition

Speech recognition was introduced into the telecommunications network in the early 1990's for two reasons, namely to reduce costs via automation of attendant functions, and to provide new revenue generating services that were previously impractical because of the associated costs of using attendants.

Examples of telecommunications services which were created to achieve cost reduction include the following:

i. *Automation of Operator Services*

Systems like the Voice Recognition Call Processing (VRCP) system introduced by AT&T or the Automated Alternate Billing System (AABS) introduced by Nortel enabled operator functions to be handled by speech recognition systems. The VRCP system handled so-called 'operator assisted' calls such as Collect, Third Party Billing, and Person-to-Person, Operator Assisted Calling and Calling Card calls. The AABS system automated the acceptance (or rejection) of billing charges for

reverse calls by recognizing simple variants of the two-word vocabulary Yes and No[18].

i. Automation of Directory Assistance

System was created for assisting operators with the task of determining telephone numbers in response to customer queries by voice. Both NYNEX and Nortel introduced a system that did front end

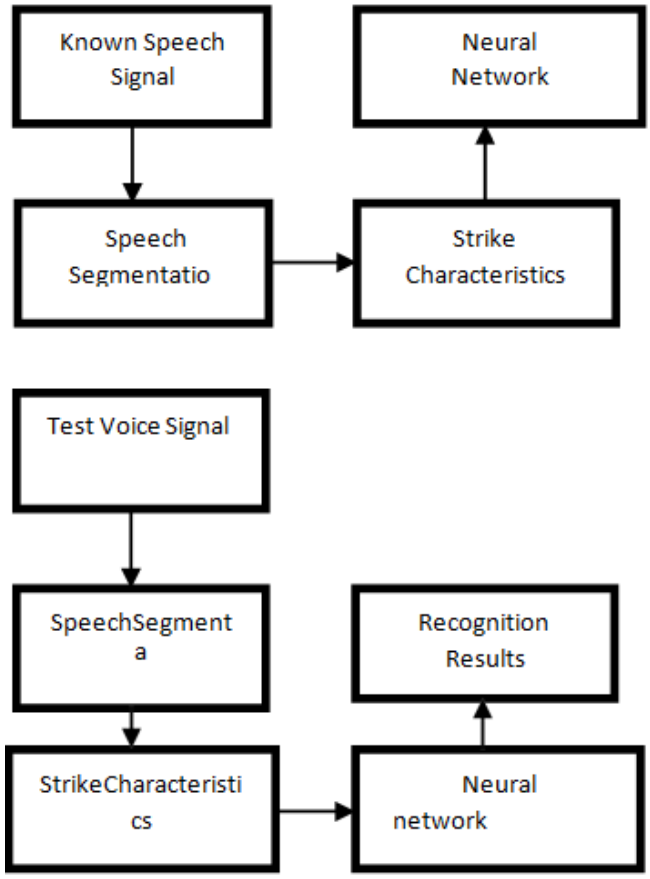

Figure 2: Neural network in speech recognition

city name recognition so as to reduce the operator search space for the desired listing, and several experimental systems were created to complete the directory assistance task by attempting to recognize individual names in a directory of as many as 1 million names[19][22]. Such systems are not yet practical (because of the confusability among names) but for small directories, such systems have been widely used (e.g., in corporate environments).

ii. *Voice Dialing*

Systems have been created for voice dialing by name (so-called alias dialing such as Call Home, Call Office) from AT&T, NYNEX[20][21], and Bell Atlantic, and by number (AT&T SDN/NRA) to enable customers to complete calls without having to push buttons associated with the telephone number being called.

VI. CONCLUSION

In this paper, given a review of Speech recognition. The area of Speech recognition is continually changing and improving. Speech recognition technology is capable to make possible to communicate with disabled persons. It makes control of digital system. In future, vast possibilities to enhance the area of speech recognition technology. By enhancing of speech recognition can provide better services for disable persons. Speech recognition can provide a secure environment to our system by voice authentication. Different methods and their accuracy also tabulated that shows the use of HMM and ANN model [19] [20] is much wider used methods for continuous speech recognition process. In the future, the correctness of speech recognition and the quality of speech will be more improve that's makes communication so easy and reliable for everybody including disable persons. Future systems must be more efficient and capable compare to traditional systems.

VII. REFERENCES

1. Jianliang Meng, Junwei Zhang and Haoquan Zhao, "Overview of the Speech Recognition Technology", 2012 Fourth International Conference on Computational and Information Sciences, 978-0- 7695-4789-3/12$26.00©2012 IEEE.
2. Andress S. Spanias, Frank H. Wu, "Speech Coding and Speech Recognition Technologies: A Review", CH3006-4/91/0000-0572$1.000 IEEE.

3. Jeff Zadeh, "Technology of speech for a computer system", DECEMBER 2003/JANUARY 2004, 0278- 6648/03/$17.00 © 2003 IEEE.
4. E Chandra and C. "A review on Speech and Speaker Authentication System using Voice Signal feature selection and Extraction", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
5. Santosh K.Gaikwad, Bharti W.Gwali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
6. Lawrence R. Rabiner, "Applications of speech recognition in the area of telecommunication", 0- 7803-3698-4/97/$10.00 0 1997 IEEE.
7. Tingyao Wu, D. Van Compernolle, H. Van hamme, "Feature Selection in Speech and Speaker Recognition" June 2009. U.D.C. 681.3_I27. Phd Thesis.
8. Urmila Shrawankar, Vilas Thakar, "Techniques for Feature Extraction in Speech Recognition System : A Comparative Study".
9. Chris Biemann, Dirk Schnelle-Walka, "Unsupervised acquisition of acoustic models for speech-to-text alignment", Master-Thesis von Benjamin Milde 10. April 2014.
10. Maxim Khalilov, J. Adri´an Rodr´iguez Fonollosa, "New Statistical And Syntactic Models For Machine Translation", TALP Research Center, Speech Processing Group, Barcelona, October 2009.
11. Richard D. Peacocke, Daryl H. Graf, "An Introduction to Speech and Speaker Recognition", Bell-Northern Research,IEEE August 1990.
12. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition", Digital Object Identifier 10.1109/ MSP.2012.2205597,Date of publication: 15 October 2012.
13. Lin-shan Lee and Yi-cheng Pan, "Voice-based Information Retrieval How far are we from the text- based information retrieval?", IEEE ASRU 2009.
14.] Masanobu Fujioka, Seiicbi Yamamoto, Naomi lnoue, Makoto Nakamura and Takashi Mukasa, "Experience and Evolution of Voice recognition applications for telecommunicati0ns services" 0- 7803-4984-9/98/$10.00 0 1998 IEEE.
15. Joseph Picone, "Continuous Speech Recognition Using Hidden Markov Models", IEEE ASSP MAGAZINE JULY 1990.
16. Todd A. Stephenson, Mathew Magimai Doss and Hervé Bourlard,"Speech Recognition with Auxiliary Information", IEEE transactions on speech and audio processing, vol. 12, no. 3, May 2004.
17. Nihat Öztürk and Ulvi Ünözkan, "Microprocessor Based Voice Recognition System Realization", 978- 1-4244-6904-8/10/$26.00 ©2010 IEEE.
18. José Leonardo Plaza-Aguilar, David Báez-López, Luis Guerrero-Ojeda and Jorge Rodríguez Asomoza, "A Voice Recognition System for Speech Impaired People", Proceedings of the 14th International Conference on Electronics, Communications and Computers (CONIELECOMP'04) 0-7695-2074- X/04 $ 20.00 © 2004 IEEE.
19. Olli Viikki, David Bye and Kari Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", 0-7803- 4428-6/98 $70.08 0 1998 IEEE.
20. Yifan Gong, "Speech recognition in noisy environments: A survey", Speech Communication 16 (199.5) 261-291,0167-6393/95/$09.50 0 1995 Elsevier Science B.V.

21. Steve Renals, Nelson Morgan, Herve Bourlard and Michael Cohen, "Connectionist Probability Estimators in HMM Speech Recognition", IEEE Transactions on Speech and Audio Processing, VOL. 2, NO. 1, PART 11, JANUARY 1994.
22. Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 22, NO. 10, OCTOBER 2014.
23. Alin G. Chit¸u, Leon J.M. Rothkrantz, Pascal Wiggers and Jacek C. Wojdel, "Comparison between different feature extraction techniques for audio-visual speech recognition", Journal on Multimodal User Interfaces, Vol. 1, No.1, March2007.

AUTHOR'S PROFILE

Mr.C.Venish raja working as a Assistant Professor in the Department of Information Technology, St.Joseph's College (Autonomous) Trichy, India.
He received his M.Phil Degree in Bharathidasan University, Trichy, India in 2012 and also He is pursuing Ph.D (Computer Science) in Bharathidasan University.

Ms.M.Daisy Jackuline is studying II M.Sc Computer Science in the Department of Information Technology, St. Joseph's College (autonomous) Trichy, India.

Ms. M.Ranjitha is studying II M.Sc Computer Science in the Department of Information Technology, St. Joseph's College (autonomous) Trichy, India