Digital Life Assistant using Automated Speech Recognition

Seema Rawat
Assistant Professor
Amity University Noida, India
Srawat1@amity.edu

Parv Gupta
B.Tech(CS&E)
Amity University Noida, India
guptaparv@amity.edu

Praveen kumar
Assistant Professor
Amity University Noida, India
pkumar3@amity.edu

***Abstract*-Physical interaction in order to provide commands or gain access to a computer system is now a history. Voice or speech stimulated systems are a part of modern Smartphone culture. Automatic Speech Recognition is an important application of artificial intelligence. This paper provides a brief description of what automatic speech recognition is, its various types and a overview of how the process works. After a precise review of Hidden Markov Model (HMM) & Mel Spectrum Cestrum Coefficient (MFCC), this paper discusses about the history of this technology, future aspects and scope. Applications of the technology in various fields are also discussed.**

Keywords- Digitized or Digital Life Assistant, Voice or Speech recognition, Hidden Markov Model (HMM), Feature extraction, MFCC, PLP, LPC.

I. Introduction*:*

All the iron man fan's are pretty much familiar with the great J.A.R.V.I.S. The J.ARV.I.S. in its movie version was an artificial intelligent based computer system that could interact with his master Tony Stark, and controls the environment of his mansion[6].

The real life JARVIS developed by Chad Barraford is a digital life assistant based application which could setup alarm clock, read out the everyday weather information and mails [6]. The system also keeps a check at who is entering the home. J.A.R.V.I.S can control the switches, lights, fans i.e. the electrical appliances. In illness mode it automatically sends a mail to the some family members or friends added in its list. All the actions said above are followed by voice command of the user. J.AR.V.I.S is modernized artificial intelligence application that makes use of automatic speech recognition technology [6].

Artificial intelligence studies the thought process of human beings and represents these thoughts by means of machines(e.g. computers, robots etc) .Natural language processing (NLP) provides a way by means of which a computer or a machine could be communicated or given instructions in a natural language that is understood and spoken by humans for e.g. English , Japanese etc. [3],[4]. Speech recognition also known by names-computer speech recognition or automatic speech recognition is the technology by means of which a computer or any machine could interpret the human voice and accomplish the said task. Speech recognition is an important application of NLP which allows physically challenged persons to provide instructions to computers without the click of buttons [4]. Speech recognition has found their important usage in military operations. It is used for important decision making process such as to command an autopilot system, weapons fire and release operations and to control the display of flight[2].

II. Typology:

2.1Speaker Dependent speech Recognition System: A speaker dependent system is one which operates upon the voice command of a single person. The computer is trained to recognize the speech and voice quality of a particular individual [3]. The voice qualities diversify from human to human. To build an artificial intelligence system that could recognize anybody's voice, is not an easy task. The system would require a very large database which is again a difficult task to accomplish [2]. Therefore, speaker dependent voice recognition systems are developed comparatively with greater ease, are economically more feasible and are simple and reliable.

2.2 Speaker Independent speech Recognition System: Even though the voice quality differs from to a great extent from one speaker to another, speaker independent voice recognition systems can be used by any individual and can perform upon any voice [5]. In these kinds of systems, user doesn't needs to train the system [3]. Therefore speaker independent voice recognition systems are less cost effective and more complicated. Also, they possess limited vocabularies.

III. Overview of speech Recognition:

The major processes of speech recognition include feature extraction, acoustic modeling, pronunciation modeling and decoder [6]. The end user gets through the application by means of an applicable input device such as a microphone. Sound waves travel in form of analog signal thus the recognizer first accepts them as analog signal and converts them into digital signal .speech signal then takes form of electrical signals [1]. Feature extraction eliminates various sources of information; it eliminates periodicity of pitch, amplitude of excitation signal and fundamental frequency etc. by combining and optimizing information, decoder performs the actual decision regard recognition of a speech utterance [7]. Figure 1 shows the block diagram of speech recognition process.

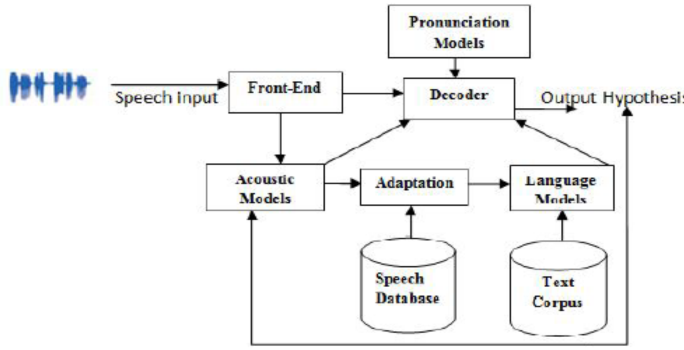

Fig 1. Outline of speech recognition system.

3.1 Feature Extraction: In automatic speech recognition system, feature extraction is the process by means of which unwanted and redundant information is discarded while retaining the useful information at the same time[11]. The process also involves transformation of signal into an appropriate form for the models used for classification. Feature extraction involves conversion of signals into digital form (known as signal conditioning), measurement of energy or frequency response of signals (known as signal measurement), signal parameterization and forming observation vectors [2]. Various techniques could be used to extract features. The most popular feature extraction technique being used is Mel Frequency Cepstrum Coefficient (MFCC) while other significant techniques are Perceptual Linear Prediction (PLP) and LPC (Linear Predictive Codes) [8].

3.1.2 Mel Frequency Cepstrum Coefficient (MFCC): The cepstrum may be defined as spectrum of a spectrum. A spectrum provides information regarding the frequency components of a signal whereas a cepstrum gives information about how those frequency changes. Framing, windowing, DFTH, Mel filter bank algorithm, computation of inverse DFT are the basic computational steps of MFCC [8]. The most common application of MFCC is their property by virtue of which system automatically recognizes the numbers spoken into a telephone; also it is used in speaker recognition which is the task of recognizing different speaker from their voices[1].

3.1.3 Perceptual Linear Prediction (PLP): PLP is based on short term spectrum of speech. In feature extraction process, PLP model represents the Physcophysics of human auditory in a much more accurate manner. The method goes vulnerable when frequency response of communication channel modifies the short term spectral values. Through the frequency band, PLP claims to avoid uniform weighting [5].

3.1.4 Linear Predictive Code (LPC): At low bit rates, linear predictive coding is assumed to be one of the most powerful speech analysis technique which provides extremely accurate parameters of speech. LPC assumes that speech signal is produced by means of a buzzer at the end of a tube (known as voiced sounds) [4].Space between the vocal folds (known as glottis) is responsible for buzz production. The tube is formed by vocal tracts (the mouth and throat). Resonance in vocal tracts produces enhanced frequency bands in sound produced. The LPC family of methods finds their uses from the basic telephony to military communications. Some sample based music synthesizers make use of LPC algorithm for compression of waveform ROM [8].

3.2 Decoding: In speech recognition process, decoding may be viewed as crucial process. It is done in order to determine the sequence of words that would match with the acoustic signal. Feature vectors are responsible for generation of the acoustic signal [6]. The essential information sources required for decoding process are –

- An acoustic model with an HMM for each unit.
- A dictionary or a list of words and the phoneme sequences they consist of,
- A language model consisting of word sequence.

3.2.1 Acoustic Modeling (Hidden Markov Model): Acoustic model is implemented using Hidden Markov Models (HMM's). HMM is a tool that is used for representation of probability distributions over a sequence of observations. On underlying principles of Markov process, HMM is a double stochastic model [7]. The semantics of the model is encapsulated in hidden part i.e. the sequence of states is hidden from observer and he could view only the output symbol sequence and therefore the model is known to be as Hidden Markov Model[7]. In the process of Automatic Speech Recognition, HMM is used to imitate a word, where phoneme represents the hidden part of the word and the statistical characteristics of corresponding acoustic events are accounted by the observable part in the given feature space. HMM can be mathematically expressed as

- A set S of N states, S = {S1, S2, S3….Sn} that are discrete and definite values. It is also assumed that hidden stochastic process could occur [9].
- An basic state probability division,

  $$\prod=\{D_L(E_j \mid t=0),E_j \in E\}, \qquad (1)$$

  where t is called as discrete time index [9].
- Another probability division which allows transition between the states, where our transition probabilities are time independent t-

  $$a_{jk}=\{D(E_k=R_{t+1}|E_j=R_t)\} \qquad (2)$$

- A set of output probability distributions that gives statistical properties of the model:

  $$c_j(x)=\{D(o_t=v_x|R_t=i)\}. \qquad (3)$$

3.2.2 Language Modeling: Language models are used for probability estimation of each hypothesized sequence of words. This probability is useful in guiding the search towards linguistically probable sequence of words as they are assigned higher probabilities as compared to unlikely sequence of words. The acoustic score is combined with language model score so as to determine how probable the hypothesized sequence of word is [5].

3.3 pronunciation modeling: A pronunciation model links acoustic model and the language model. This model compares the words generated by acoustic model with the words present in dictionary and produce's string of words which is our final output. It contains information about which words are known to system or how they may be pronounced (known as phonetic representation) [2].

IV. Historical Perspectives of Voice Recognition:

Although first speech recognizer was developed in 1952 at Bell Laboratories, the speech recognition system for Pc was developed in 80's. Speech dictation system came into existence around 1990's.To handle the continuous speech, first dictation software was developed at Dragon Systems by Jim and Janet baker in the year 1997 [11].

V. Speech Recognition Forges Ahead:

Human beings have been able to recognize the voice of other human's since ages when languages just began. Soon an era would arrive when we could speak directly to our mobiles, tablets and laptop and could instruct them. Then it would send a mail on your voice commands or would format what you said to what you need-a report, a letter, a memo or whatever. Currently three research teams of Defense Advanced Research Projects Agency (DARPA) are functioning on Global Autonomous Language Exploitation (GALE), this program is alleged to accept streams of information from newscast and newspapers of foreign countries and translate them. DARPA claims that the software could quickly translate two languages with an expected accuracy of 90 percent [10].Sony, IBM, Olympus, Lernout, Norcom, Dragon Systems, Phillips, and Grover Industries are a few names that are following the thrill of voice recognition by developing software and portable devices with higher feats of accuracy.

VI. Battle of Digital Life Assistants:

Without a footnote of Apple's Siri, Google Now, Microsoft Cortana and Samsung's S Voice discussion on digital life assistant could never end. They are voice respondent systems for mobile applications. These systems make use of traditional Natural Language Processing and Speech Recognition System [6]. However if State System could be applied to these applications response time and user experience could be enhanced to much greater extent.

The culture was started by Apple's Siri for ios devices, Google Now soon accompanied next with almost same ingredient for android platform and Microsoft Cortana has just joined the race for windows 8.1 Phone users [11]. All these services are web based-speech recognition gets invoked as soon we talk to the application. The words are then parsed in the form of queries which are sent to some web based service and answer is displayed or spoken out by the device.

VII. Conclusion:

Speech Recognition is an excellent application of artificial intelligence to work with since speech is one of the most basic activities of humans. Through this paper an attempt has been made to provide a review of how this technology actually works and how far the progress had been made in past few years. Computers would quickly arrive with preinstalled automatic speech recognition systems. Equipments and devices with this technology would make the lives of the blind, the deaf, and other physically challenged people by providing them access to computers without the click of buttons. Voice recognition would soon be a feature of every small to big devices ranging from ticket selling devices to washing machines e.tc.

References:

[1] Marcus E.Hennecke , K.Venaktesh Prasad, David G. Stork , "Automatic Speech Recognition System using Acoustic & Visual Signals" , 1996 IEEE Proceedings of ASILOMAR-29.

[2] Preeti Saini , Parneet Kaur, "Automatic Speech Recognition" , international journal of engineering tools and technology-volume 4 issue 2-2013.

[3] Raghvendra Priyam, Rashmi Thakur, Videh Kishori Thakur, "Artificial Intelligence Applications for Speech Recognition", CAC2S 2013.

[4] Gruhn, RE; Minker,W;, Nakamura S. ,"Statistical Pronunciation Modelling for Non-Native Speech Processing" ,ISBN: 978-3-642-19585-3

[5] Shrutika Khobragade, "Jarvis Digital Life Assistant", international journal of engineering research and technology-volume 2 issue 1-2013.

[6] Kishore Venkateshan, "Switch State Mechanism for Digital Life Assistant", 3rd international conference on intelligent computational systems-ICICS 2013, Singapore.

[7] R.E Gruhn et al., "statistical pronunciation modeling for non-native speech processing", Springer Verlag Berlin Heidelberg-2011.

[8] Nmarata Dave, "Feature Extraction Methods, LPC, MFCC in speech recognition, international journal of engineering research and technology-volume 1 issue VI-2013.

[9] Xian Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for ASR" , 2009 Pacific Asia Conference on Circuits, Communication & Systems.

[10] http://www.ieeexplore.ieee.org

[11] http://www.google.com.