Voice Controlled Personal Assistant

[1]U. Rajesh Naidu, [2]M. Jaya Laxmi, [3]V. Sai Sumanth, [4]A. Nagamani, [5]V.V.K. Raju
Department of Electronics and Communication Engineering,
Anil Neerukonda Institute of Technology and Sciences
Andhra pradesh, India – 531162

***Abstract* — Speech recognition technology is one of the fast growing technologies. Nearly 20% people of the world are suffering from various disabilities; many of them are blind or unable to use their hands effectively. These systems in those particular cases provide a significant help to them, so that they can share information with people by operating computer through voice input. Hidden Markov Model is one of the algorithm is used to convert speech into text. This paper describes the development of speech recognition technology and its basic principles, methods, reviewed the classification of speech recognition systems and voice assistant technology.**

Keywords—Speech Recognition, Hidden Markov Model, Phonetics Foundation, Online Information Services, IOT.

I. INTRODUCTION

Speech recognition is the program is to identify the human voice, understand and react accordingly to it. It is based on the voice as control input, it allows the machine to automatically identify and understand human spoken language through speech signal processing and pattern recognition. This technology allows the machine to convert speech signal into text. It is a cross-disciplinary and involves a wide range. It has a very close relationship with acoustics, phonetics, linguistics, information theory, pattern recognition disciplines. With the rapid development of computer hardware and software technology, speech recognition is gradually becoming a key technology. Products to develop speech recognition technology is also widely used in voice activated telephone exchange query information networks, medical services, banking services, industrial control every aspect of society and people's lives.

II. THE DEVELOPMENT PROCESS

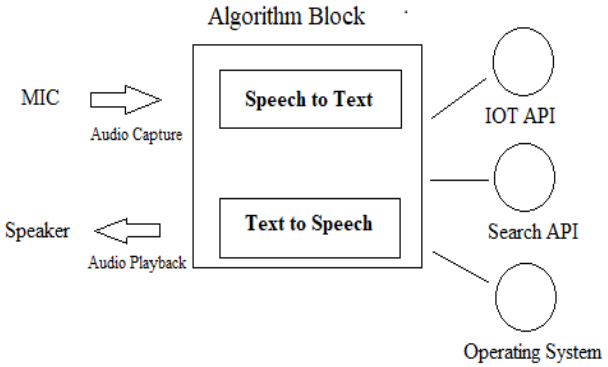
Speech recognition research work began in the 50's, Bell Labs speech recognition system-Audrey system first identifies the ten English digits. But it really made substantial progress, and as an important issue in conducting research in the late 60's the early 1970s. Further speech recognition in the 1980s, the HMM model and artificial neural network (ANN) are successfully used in speech recognition. 1988, FULEE Kai and others use the VQ/IIMM method to achieve speaker-independent continuous speech recognition system-SPHINX, including 997 vocabulary. This is the first of the world speech recognition system, it is a high-performance, non-specific, large vocabulary continuous speech recognition system. People finally breakthrough of the three major obstacles, including a large vocabulary, continuous speech and non-specific. And it identified the mainstream of statistical methods and models in speech recognition and language processing. Today, Many companies like Google, Amazon and Apple are trying to achieve this in generalized form.

III. METHODOLOGY

A. *System Architecture*

The overall system design consists of following phases:

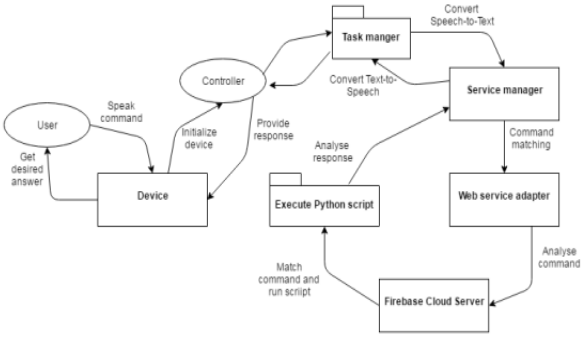
- Data collection in the form of speech.
- Voice analysis and conversion to text.
- Data storage and processing.
- Generating speech from the processed text output.



Systematic Architecture of Voice Controlled Device

B. *Data Flow Sequence*

- Initialize Device: Initializes the device by calling its name.
- Task Manager: Conversion of speech-to-Text and Text-to-Speech is performed by the task manager.
- Service Manager: It performs the analysis of commands and matching them with web service adapter and cloud server.
- Firebase Cloud Server: The data generated is stored in the firebase cloud server and the data is accessible by the main system. The data can be retrieved and processed as per requirement.



Dataflow Diagram

IV. PHONETICS FOUNDATION

Phonetics is a branch of linguistics that studies the sounds of human speech, or in case of sign languages-the equivalent aspects of sign. Human speech is produced by the vibration of the vocal cords and the configuration of the vocal tract that is composed of articulatory organs, including the nasal cavity, tongue, teeth, velum and the lips. Because some of these articulators are visible, there is an inherent relationship between acoustic and visual forms of speech. The main assumption underlying speech-driven facial animation is that the visual component of speech is more or less redundant. That is, it is presumed that sufficient portions of the visual information can be inferred from the acoustic information.

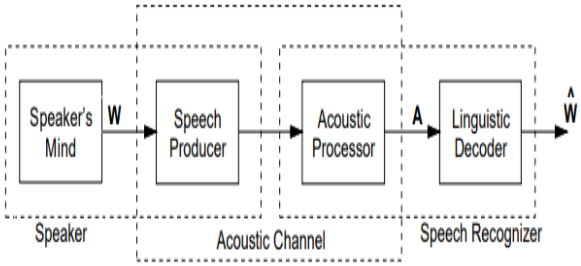
aa	car	*ey*	make	*r*	ray
ae	hat	*f*	far	*s*	sea
ah	cut	*g*	agree	*sh*	she
ao	score	*hh*	hay	*t*	tea
aw	house	*ih*	bit	*th*	think
ax	about	*ix*	accident	*uh*	book
ay	pie	*iy*	sea	*uw*	boot
b	bed	*jh*	joke	*v*	voice
ch	choke	*k*	key	*w*	way
d	day	*l*	lay	*y*	yard
dh	that	*m*	mill	*z*	zone
dx	dirty	*n*	nine	*zh*	measure
eh	get	*ng*	sing	*cl*	<closure>
el	bottle	*ow*	boat	*vcl*	<closure>
en	button	*oy*	boy	*epi*	<silence>
er	bird	*p*	put	*sil*	<silence>

48-Phoneme set with example pronunciations

V. SPEECH RECOGNITION

Speech recognition is the process of transcribing acoustic speech into text or more generally, into a sequence of labels. That talent comes so naturally for us humans that the difficulty of endowing a computer with an equal capability has repeatedly been underestimated. From the technology perspective, speech recognition has a long history with several waves of major innovations.

Speech recognition systems were built in the 1950's for vowel recognition and digit recognition, yielding creditable performance. It was thought that these results could be extended in a natural way to more sophisticated systems. Unfortunately, the techniques did not scale up a situation all too familiar from many other frontiers of artificial intelligence. The real world with its immense proportions and diversity proved impossible to handle for techniques that were developed for simple, constrained tasks. Some speech recognition systems require training where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses the voice to fine-tune the recognition of that person's speech which in an increased accuracy. Systems that do not use training are called "speaker independent". Systems that use training are called "speaker dependent".



The Source – Channel model of Speech Recognition

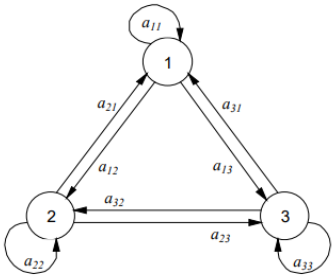
VI. HIDDEN MARKOV MODEL

As a statistical model,Hidden Markov Models Hidden Markov Model (HMM) analysis founded in the 1970s and 1980s has been the dissemination and development and successfully applied to the modeling of the acoustic signal.To the 1990s, HMM has also been the introduction of the computer word recognition and mobile communication core technology of multi-user detection. So far, it is still considered to be the most successful approach to achieve fast and accurate speech recognition system. The HMM model parameters represent the time-varying characteristics of the voice signal. It consists of two interrelated stochastic processes common to describe the statistical characteristics of the signal. One of which is hidden (unobserved) finite-state Markov chain, and the other is the observation vector associated with each state of the Markov chain stochastic process (observable). Reveal characteristics of the hidden Markov chain depends on the signal characteristics can be observed. In this way, a certain period of time varying signals such as voice characteristics described by the random process corresponding to the symbols of state observation. Signal described by the hidden Markov chain transition probability changes with time.

A. Markov Chains:

Let $\{S_0 , S_1 , ..., S_t, ...\}$ be a sequence of discrete random variables assuming values in a finite alphabet 6 = {1, 2, ..., *N*}. The random variables are said to form a *Markov chain*, if for all values of *i* greater than zero.

$$P(S_i = s_i \mid S_0 = s_0, K, S_{i-1} = s_{i-1}) = P(S_i = s_i \mid S_{i-1} = s_{i-1})$$



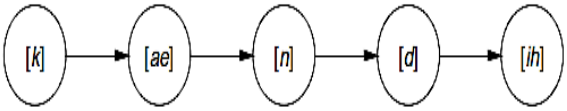
A Markov Model with Three States

B. Hidden Markov Model in Speech Recognition

Hidden Markov models have enjoyed widespread popularity in speech recognition research for the past two decades. The success of HMMs in speech processing is mostly explained by three factors:

- They have a sound mathematical basis and practical algorithms for training and use;
- They are good at modeling one-dimensional time-varying signals in general;
- They are good at modeling uncertainty and do not require too many assumptions to be made about the source or process that is to be modeled.

In continuous-speech recognition, *subword models* are required to account for differences in pronunciation and for the possibly infinite vocabulary. The subword models are then used as building blocks for whole-word models. For example, an acoustic model for the word *candy* can be constructed by concatenating the five phoneme models for *k*, *ae*, *n*, *d* and *ih*, or the two syllable models for *kaen* and *dih*. This is illustrated in Figure.



A word HMM composed by concatenating phoneme HMMs. Each state in the world model is actually an HMM representing a specific phoneme.

VII. Application of Speech Recognition Technology and the Facing Problems

A. Application of Speech recognition technology

The world to speed up research and development of speech recognition applications, there are some practical speech recognition system put into commercial operation. The typical speech recognition system-VRCP system developed by AT&T in 1992 .The system is five words (collect, person, third number, the operator and calling card), non-specific small-vocabulary speech recognition system, has been used in AT&T Communications online, you can achieve the automatic operator-assisted call, instead of the operator completed five kinds of call type. In September 1996, Charles Schwab launched the first large-scale commercial speech recognition application systems: the stock quotation system. The system was also the first in the financial field speech recognition system. The system is effective to improve the quality of service and customer satisfaction, and reduce call center costs. Soon, Schwab opened the speech of stock trading system. Departments in major U.S. telecom operator Sprint PCS has the largest digital wireless network at the same time' known for excellence and innovative customer service. The opening voice-driven systems for clients since 2000. The system provides customer service, voice dialing, check number, and change addresses and other services. In addition, China Telecom has launched a voice recognition integration of value-added services system CELL-VVAS, (VOICEVALUE-ADDED SYSTEM), the system uses a distributed excellent recognition engine, developed a stable and efficient application. The system also perfectly integrated telecommunications switching network application to provide users with a variety of user-friendly, personalized service [7]. Another development branch of speech recognition technology is the development of the telephone voice recognition technology, Bell Labs is a pioneer in this regard, the telephone voice recognition technology will be able to telephone inquiries, automatic wiring, as well as some specialized operations, such as tourist information and other operations. After the bank use the voice query system of speech understanding technology, it can provide customers with 24-hour Phone Banking Service. Securities industry, using telephone speech recognition audio system, then, the user would like to query market could speak out the stock name or code system to confirm the user's requirements, will automatically read the latest stock price, which will greatly facilitate the user . In the 114 directory assistance artificial voice technology, you can let the computer to automatically answer the needs of users, and then playback the phone number of the query, thus saving human resources.

B. The facing problems

At present, speech recognition research progress has been slow, mainly in theory has been no breakthrough. Although a variety of new amendments continue to emerge, but also the lack of general applicability. Mainly in: Poor adaptability of the speech recognition system is mainly reflected in the dependence on the environment, If you collected speech training system in certain circumstances, the system can only be application in this environment, otherwise the system performance will be a sharp decline, another problem is that this system does not respond correctly for the error input of users. Additionally, the progress of speech recognition in noisy environments is very difficult, because at this time people's pronounce varies greatly, like voice, slow speech rate, pitch and format changes, which is the Lombard effect, must find a new signal analysis and processing approach.

Understanding of the human auditory comprehension, and accumulation of knowledge and learning mechanism and system of the brain control mechanism is still unclear, and secondly, the existing achievements of this aspect is used in speech recognition also remains a difficult process.

VIII. Conclusion

Speech recognition has a big potential in becoming an important factor of interaction between human and machine in the near future., speech recognition systems are widely in several applications . However, in near future the voice recognition technology continues to improve, the speech recognition system will be more in-depth, the application of

speech recognition systems will be more extensively used [8] and a variety of speech recognition systems will appear in the market and people will adjust their speech patterns to adapt to a variety of recognition system. In future the usage of voice assistants will be more , we can forward step by step direction to improve the speech recognition system. In future the speech recognition systems will play a major role.

REFERENCES

[1] Yu Tiecheng. The current development of speech recognition [J]. Communication World, 2005.
[2] Ren Tianping. Application of speech recognition technology [J]. Henan Science and Technology, 2005.
[3]L A Liporace. Maximum Likelihood for Multivariate Observation of MarkovSources. IEEE.Trans. IT, 1982, 28(5): 729-734.
[4] Zhang Ping, Zhang Qiong. Based on HMM and BP neural network for speech recognition [J]. Cross-century, 2008.
[5] Yin Peng, Li Tao, Wang Haibing. Intelligent neural network system composed of the principle in speech recognition. Mini-Micro Systems,2000,21(8):836-839.
[6] Jiang Ming Hu, in the Yuan Baozong, Lin Biqin. Neural networks for speech recognition research and progress. Telecommunications Science,1997,13(7):1-6.
[7] Huang Shan. Voice recognition systems in the telecom prepaid business applications [J]. Information Science, 2010.
[8] Yangshang Guo, Yang Jinlong. The speech recognition technology overview [J]. Computer, 2006.