Speech Recognition : A Comprehensive Study

Mohsin Manshad Abbasi, Dr. Abdul Majid Abbasi, Anees Qumar Abbasi

Abstract— Speech recognition is a topic of modern research. It is the process of converting spoken input into text. Different techniques are used to design equipment's that are used for speech recognition. In this study, different methodologies, techniques, hard ware and software are discussed in a precise manner. In last section, future expectations are discussed in a detailed manner.

Index Terms— Dependent Systems, Grammer of Speech, Independent Systems, Loudness in Speech, Speech Recognition Methodolgies, Types, Utterance..

1 INTRODUCTION

Speech recognition is the process of converting spoken in put into text. It is also known as Automatic Machine Organization or Speech To Text (STT) and Voice Recognition (VC).Automatic Speech Recognition is the process by which a computer maps an audio speech signal to text. In this process, a computer maps an audio speech signal to some form of abstract meaning of the speech.

Speech Recognition use speaker independent SR.This system analyzes person's specific voice and uses it to fine tune the recognition of that person's speech resulting in more accurate transcription.Those systems that do not use training are called speaker independent system while those systems that use training are called speaker dependent systems.

SRA (Speech Recognition Application) includes voice user interface such as voice dialing (e.g. "Call home"), domestic-control, and search,and simple data entry, preparation of structured documents, speech-to-text processing and aircraft.

2. BASICS OF SPEECH RECOGNITION

Following are a few of the basic terms and concepts that are essential for speech recognition. It is important to have a good understanding of these concepts.

- Utterances
- Pronunciation
- Accuracy
- Grammar
- Speaker dependent systems
- Speaker independent systems
- Training

2.1 Utterance:

It is defined as once the user says something; this is known as an **utterance.** An utterance is any brook of speech between two periods of silence. They are sent to the speech engine to be handled. Quietness, in speech recognition, is almost as important as what is spoken, because quietnessdefines the start and end of an utterance.The Speech Recognition Engine is "attending" for speech input, whenever the engine notices an audio input - in other words, a lack of silence. Correspondingly, once the engine notices a certain amount of silence following the audio, the end of the utterance arises.

Utterances are referred to the speech engine to be handled. If the user does not say something, then the engine returns. Asuggestion that there was no speech perceived within the expected timeframe and the application takes a suitable action, such as motivating the user for an input. An utterance can be a single term, or can be a phrase or a sentence.

2.2 Pronunciation:

The speech recognition engine uses different types of data, statistical models, and algorithms to translatevocal input into text. The information which is used by speech recognition engine is the pronunciation of word which shows that what the speech engine thinks that a word should sound like. Many words have different pronunciation related with them. For example, the word "the" has at least two pronunciations in the U.S. English language: "thee" and "thuh." For Voice XML application different pronunciations are used for certain words or phrases to get the results which helps to recognize the caller voice.

.

2.3 Accuracy:

The act of a speech recognition system is measurable. Possibly the most generally used measurement is accuracy. It is usually a quantitative measurement and can be calculated in several

- *Mohsin Manshad Abbasi is currently pursuing doctorate Degree in Computer Sciences fromDepartment of Computer Sciences & Information Technology, University of Azad Jammu & Kashmir, Pakistan.. E-mail: mohsinmanshad@gmail.com.*
- *Dr.Abdul Majid Abbasi is currently working as Assistant Professor in Department of Computer Sciences & Information Technology, University of Azad Jammu & Kashimr, Pakistan*
- *Anees Qumar Abbasi is currently pursuing doctorate Degree in Computer Sciences fromDepartment of Computer Sciences & Information Technology, University of Azad Jammu & Kashmir, Pakistan.*

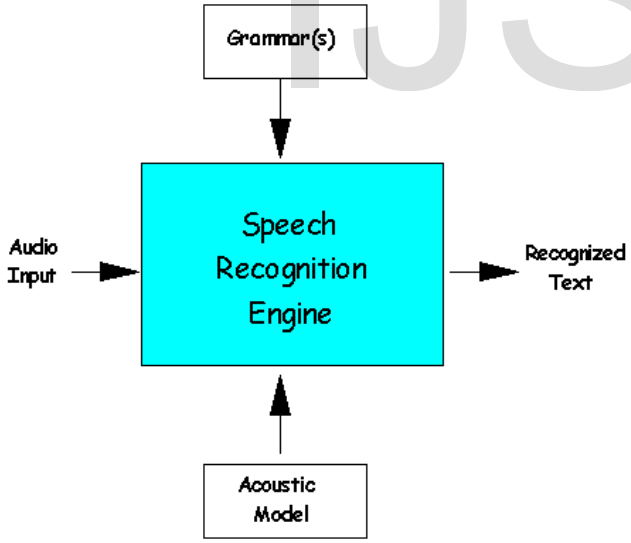
ways. Arguably the most important measurement of accuracy is whether the preferredend result occurred. This measurement is useful in authenticating application design. For example, if the user said "yes," the engine repaid "yes," and the "YES" action was executed, it is clear that the preferred end result was achieved. But what happens if the engine repaid text that does not exactly match the utterance? For example, what if the user said "nope," the engine repaid "no," however the "NO" action was executed? Should that be considered a successful paper? The answer to that question is yes because the preferred end result was achieved.

How it works

Nowadays that we have discussed some basic terms and concepts that included in speech recognition, let's put them together and take a look at how the speech recognition (SR) process works.

The speech recognition engine has a rather difficult task to handle, that of taking raw auditory input and translating it to recognized text that an application recognizes as shown in the following diagram, we discussed the major components:

- Audio input
- Grammar's
- Acoustic Model
- Recognized text



2.4 Grammar:

The system requires the words and phrases. These words and phrases are appropriate to the speech recognition engine and are used in the recognition process.

A grammar uses a particular composition, or set of rules, to describe the words and phrases that can be recognized by the engine. A grammar can be as simple as a list of words. Grammars describe the domain, or context, in which the recognition engine works. The engine associates the current utterance against the words and phrases in the dynamic grammars. If the user says something that is not in the grammar, so the speech engine will not be able to decode it correctly.Grammar is basically a noun.The grammar to support this interaction strength contains the following words and phrases:

- Accounts
- account balances
- My account information
- Loans
- Loan balances
- My loan information
- Transfers
- Exit
- Help

2.5 Speaker Dependent System:

Speaker dependence defines the degree to which a speech recognition system requires knowledge of a speaker's separate voice characteristics to successfully process speech. The speech recognition engine can "acquire" how you speak words and phrases; it can be trained to your voice.

Speech Recognition Systems (SRS) that need a user to train the system to his/her voice are identified as speaker-dependent systems. If you are aware with desktop dictation systems, most are speaker dependent. Since they operate on very large words, dictation systems perform much better when the speaker has spent the time to train the system to his/her voice.A speaker dependent system is established to function for a single speaker. These systems are usually easier to grow, cheaper to buy and more precise, but not as flexible as speaker independent systems.

2.6 Speaker Independent System:

Speech Recognition Systems that do not need a user to train the system are known as speaker-independent systems. SR in the VoiceXML world must be speaker-independent. Reason of how many users (hundreds, maybe thousands) may be calling into your web site. You can't need that each caller trains the system to his or her voice. The speech recognition system in a voice-allowed web application NECESSITY successfully process the speech of many different callers without having to know the individual voice characteristics of each caller.A speaker independent system is established to operate for any speaker of a specific type (e.gAmerican English). These systems are the most difficult to develop, most exclusive and accurateness is inferior to speaker dependent systems. Though, they are more stretchy.

2.7 Training:

Training is the process concluded which employees are made skillful of doing the job arranged to them. Allowing to Flippo:"Training is the performance increasing the knowledge and skill of an employee for doing a specific job." According to Dale Yoder, "Training is the procedure by which man-power is filled for the specific jobs it is to perform." According to Beach: "Training is the procedure by which people learn knowledge and skills for a certain purpose."

3. HARDWARE

The hardware recommended for developing a Speech Recognition System are;

- Sound cards:

  Sound card with the cleanest Analog to digital conversion are recommended.
- Microphone:

  The best high-quality for microphone is the receiver style.
- CPUs or processors:

  The more the speed the well SR would work.For good SR there should be at least 1GHz processor and1GB of RAM.

4. USES/ APPLICATIONS

4.1 Military
- Helicopters
- High performance aircrafts
- Training Air Traffic Controllers(TATC)

4.2 People with Disabilities:
- Speech recognition technology (SRT) help people with disabilities interact with computers more easily.
- People with motor limitations, which can't use standard keyboard and mouse and use their voices to cross the computer and then creat, the documents.

4.3 Dyslexic People:
- SRT (speech recognition technology) is very cooperative for people with learning disabilities, who experience conflict with spelling and writing.
-

5. FUTURE APPLICATIONS:

5.1 Home Automation

There is a portion of it currently in use. In future the home appliances such as oven, washing machine, electricity supplierswill be used and controlled by Speech Recognition Systems

5.2 Wearable Computers

The most innovative application is in the use and functionality of wearable computers.

6. SPEECH RECOGNITION SOFTWARE

6.1 IBM via voice:

International Business Machines (IBM) is by far the world's largest information technology company in terms of income ($88 billion in 2000).IBM (International Business Machine) products include hardware and software for a line of business waiters, storage products, custom-designed microprocessors, and application software.

IBM via voice is the only range of language.IBM via voice is only program that will run on older or less powerful computers via voice has great dictation ability that it converts your speech into the text and then performs actions.

6.2 Microsoft Speech Recognition System:

Speech recognition system of technology are used in some Microsoftproducts, Microsoft office, office2003, office2007, Microsoft plus for window xp, windows mobiles etc. However proceeding to windows vista, SR was not main stream.
Microsoft windows to proposal fully unified support for SR.

6.3 Philips speech Magic

Speech magicis an industrial grade platformfor capturing information in a digital format.it has been developed by Philips speech recognition systems of Vienna, Austria.it technology is mainly for healthcare sector. However applications can also be availablefor the legal market.Other SR software are bable technologies, Speech works etc.

7. CONCLUSION:

Speech recognition (SR) willefficient the way people conduct business over the web and will finallydifferentiate world classe business.SR will change the technique people conduct business over the Web. VoiceXML drawsSR and provides the technology in which businesses can improve and organize.. Speech recognition (SR) and VoiceXML obviously is most important in over days.

REFERENCES

[1] Baker, J.; Li Deng; Glass, J. Khudanpur, S., Chin-hui Lee, Morgan, N, O'Shaughnessy, D., "*Developments and directions in speech recognition and understanding, Part 1 [DSP Education],*" Signal Processing Magazine, IEEE , vol.26, no.3, pp.75-80, May 2009

[2] K.-F. Lee, *Automatic Speech Recognition: The Development of the Sphinx Recognition System*. New York: Springer-Verlag, 1988.

[3] C.-H. Lee, F. Soong, and K. Paliwal, Eds.,*Automatic Speech and Speaker Recognition- Advanced Topics*. Norwell, MA: Kluwer, 1996.

[4] C. Leggetter and P. Woodland, "*Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,"Comput. Speech Lang*., vol. 9, no. 2, pp. 171–185, 1995.

[5] R. Lippman, "*An introduction to computing with neural nets," IEEE ASSP Mag*., vol. 4, no. 2, pp.4–22, Apr. 1987.

[6] H. Ney, The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 32, no. 2, pp. 263–271, 1984.

[7] Mohsin M A., " Clustering: An Effective Methodology to Identify Rare Cases In Pain" LAMBERT Academic Publishing, ISBN 978-3846510896, 2011

[8] Mohsin M A., "CDSSs: Review on different methodologies used in Health Care" LAMBERT Academic Publishing, ISBN 978-3848402120, 2012

[9] D. Paul, "*Algorithms for an optimal A* search and linearizing the search in a stack decoder,"* in *Proc. IEEE ICASSP*, 1991, vol. 1, pp. 693–696.

[10] H. Poor, *An Introduction to Signal Detection and Estimation* (Springer Texts in Electrical Engineering), J. Thomas, Ed. New York: Springer-Verlag, 1988.

[11] D. Povey, B, Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, *"FMPE: Discriminatively trained features for speech recognition,"* in *Proc.IEEE ICASSP*, Philadelphia, PA, 2005, pp. 961– 964.