ISSN: 2349-7637 (Online)
Volume-1, Issue-2, September 2014
RESEARCH HUB – International Multidisciplinary Research Journal
Research Paper
Available online at: www.rhimrj.com

Artificial Intelligence for Speech Recognition

Prof. Manish G. Gohil
Lecturer,
Computer Science
Smt. S. J. Varmora BBA & BCA Mahila College
Wadhwan City, Surendranagar
Gujarat (India)
manish.mca31@gmail.com

Abstract: ***This paper presents how Speech Recognition, the most important application of Artificial Intelligence grows in technology. The notion of AI is well known due to its popularity in science fiction movies which depict humans interacting with machines as they would with other humans. Speech and gestures are the natural means of communication used by humans to interact with each other. Speech Recognition makes it possible for you to speak to a computer. Speech Recognition Software is the technology that transforms spoken words into alphanumeric text and navigational commands. Speech Recognition is used in legal and medical transcription, the generation of subtitles for live sports and current affairs programs on television. In naturally spoken language, there are no pauses between words, so it is difficult for a computer to decide where word boundaries lie. Automatic speech recognition is the process by which a computer maps an acoustic speech signal to text. These powerful trends will drive the next generation of information technology into the mainstream by about 2010.***

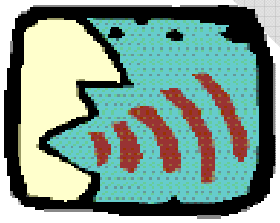
Key Words: artificial, humans, speech, affairs, information, mainstream

INTRODUCTION

Making machines more like human beings has always been a strong desire of man. This desire has led to the emergence of disciplines like artificial intelligence (AI) which emulate human behavior in machines. The notion of AI is well known due to its popularity in science fiction movies which depict humans interacting with machines as they would with other humans. The concepts used in AI include the principles outlined by man machine interfacing (MMI) which allows the creation of machines that are more usable for humans. Speech and gestures are the natural means of communication used by humans to interact with each other.

Speech technology is currently at a stage where it can be used but only in a constrained way which, although does not provide seamless interaction, does mean a step towards it. Speech recognition technology is required by a machine to be able to interpret human speech. Although speech recognition technology has been under development for many years it had not been established enough to be used with PCs until recently. Accuracy and speed are two major factors that are necessary to make speech interfaces practical for frequent use. Hardware used in personal computers is now advanced enough to supply enough processing power to be able to run speech recognition at a usable speed. Accuracy of speech recognizers is also improving. Some commercial speech recognizers can now handle continuous speech with an accuracy of more than 90%. Speech synthesis is required to allow computers to communicate back to the user in speech. Speech synthesis tools are also now widely available; examples of such tools are Microsoft's Speech API and Speech Works Speechify.

HOW TO RECOGNIZE SPEECH?



Simple inquiries about bank balance, movie schedules, and phone call transfers can already be handled by telephone-speech recognizers. Voice activated data entry is particularly useful in medical or darkroom applications, where hands and eyes are unavailable. Speech could be used to provide more accessibility for the handicapped (wheelchairs, robotic aids, etc.) and to create high-tech amenities (intelligent houses, cars, etc.)The 1990s shows the first commercialization of spoken language understanding systems. Computers can now understand and react to humans speaking in a natural manner in ordinary languages within a limited domain.

WHAT IS SPEECH RECOGNITION?

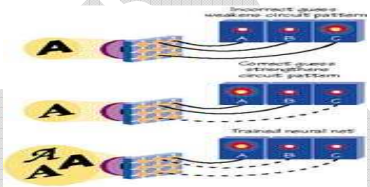
You hear a version of it every day. Each time you call your bank for balances, or call the local theater for movie times and pricing. Each time you call a business and you are directed by an automated voice you are interacting with a type of Speech Recognition, better defined as telephone-speech recognizers. Speech Recognition makes it possible for you to speak to a computer.

WHAT IS SPEECH RECOGNITION SOFTWARE?

Speech Recognition Software is the technology that transforms spoken words into alphanumeric text and navigational commands. To increase dictation precision, it generates an additional dictionary of the words used. A main factor of Speech Recognition Software is the language model.

WHAT IS THE LANGUAGE MODEL?

Speech Interfacing

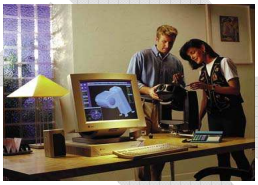


Recognition of Speech Frequency

The Language Model analyzes your speech. In theory the Language Model decides which words you said and shows them on the screen. A new technology, called natural language speech recognition, is markedly improving voice-activated self-service. Powered by artificial intelligence, these speech-recognition systems are altering consumer perceptions about phone self-service, as calls for help no longer elicit calls for help.

WHAT DO YOU NEED TO RUN SPEECH RECOGNITION SOFTWARE ON YOUR COMPUTER?

To run Speech Recognition Software on your computer, you will need a fast CPU at least a minimum of 400MHz, and plenty of RAM at least 128M. The next item you will need will be a good quality microphone. This is the key when using Speech Recognition. Desktop microphones will not do the job; they tend to pick up more ambient noises. Hand held microphones are not a good idea; they can be cumbersome to pick up all the time and they do not limit the amount of ambient noise. The best choice is the headset style microphones. The ambient noise is minimized and the microphone stays close to your mouth all the time. The last item you will need is a good sound card.



Camera

Microphone (Voice)

Translator

SOFTWARE DESIGNED TO MAKE LIFE EASIER

With Speech Recognition software you can write emails, memos, reports, list anything you would normally type on the computer. With Speech Recognition you can tell your computer what to do. With Imagine opening web pages with just your voice or dialing phone numbers from your address book. The possibilities are endless.

WHAT SHOULD YOU DO BEFORE PURCHASE SPEECH RECOGNITION SOFTWARE?

- Compare prices and shop for the best deals
- Determine how many languages can the software support
- Determine what type of technical support is offered
- Make sure your computer has plenty of Ram, a good microphone, and a good sound card.

ISSN: 2349-7637 (Online)

CLASSIFICATION OF SPEECH RECOGNITION

- Whether they require the user to "train" the system to recognize their own particular speech patterns or not.
- Whether the system is trained for one user only or is speaker independent.
- Whether the system can recognize continuous speech or requires users to break up their speech into discrete words.
- Whether the system is intended for clear speech material, or is designed to operate on distorted transfer channels (e.g. cellular telephones) and possibly background noise or another speaker talking simultaneously.
- The context of recognition - digits, names, free sentences.

USE OF AI FOR SPEECH RECOGNITION

Despite the apparent success of the technology, few people use such speech recognition systems on their desktop computers. A typical office environment, with high amplitude of background speech, is one of the most adverse environments for current speech recognition technologies, and large-vocabulary systems with speaker-independence that are designed to operate within these adverse environments have significantly lower recognition accuracy. The typical achievable recognition rate as of 2005 for large-vocabulary speaker-independent systems is about 80%-90% for a clear environment, but can be as low as 50% for scenarios like cellular phone with background noise.

Speech recognition systems have found use where the speed of text input is required to be extremely fast. They are used in legal and medical transcription, the generation of subtitles for live sports and current affairs programs on television; not directly but via an operator that re-speaks the dialog into software trained in the operator's voice; in such cases the operator also has special training, first to speak clearly and consistently to maximize recognition accuracy, second to indicate punctuation by various techniques, and also often domain-specific training (especially in medical or legal contexts). In courtrooms and similar situations where the operator's voice would disturb the proceedings, he or she may sit in a soundproofed booth or wear a Steno mask or similar device.

Speech recognition is sometimes a necessity for people who have difficulty interacting with their computers through a keyboard, for example, those with serious carpal tunnel syndrome, damaged hands or arms, or other physical limitations. Speech recognition technology is used more and more for telephone applications like travel booking and information, financial account information, customer service call routing, and directory assistance. Research and development in speech recognition technology has continued to grow as the cost for implementing such voice-activated systems has dropped and the usefulness and efficacy of these systems has improved. For example, recognition systems optimized for telephone applications can often supply information about the confidence of a particular recognition, and if the confidence is low, it can trigger the application to prompt callers to confirm or repeat their request. Furthermore, speech recognition has enabled the automation of certain applications that are not automatable using push-button interactive voice response (IVR) systems, like directory assistance and systems that allow callers to "dial" by speaking names listed in an electronic phone book. Nevertheless, speech recognition based systems remain the exception because push-button systems are still much cheaper to implement and operate. Speech recognition is also used for speech fluency evaluation and language instruction.

The application of computer speech recognition, though more limited in utilization and practical convenience, has made it possible to interact with computers by using speech instead of writing. Modern speech recognition systems are generally based on hidden Markov models (HMMs). This is a statistical model which outputs a sequence of symbols or quantities. Having a model which gives us the probability of an observed sequence of acoustic data given one or another word (or word sequence) will enable us to work out the most likely word sequence by the application of Bayes' rule:

$$\Pr(\text{word} \mid \text{acoustics}) = \frac{\Pr(\text{acoustics} \mid \text{word})\Pr(\text{word})}{\Pr(\text{acoustics})}.$$

For a given sequence of acoustic data (think Wave file), Pr(acoustics) is a constant and can be ignored. Pr(word) is the prior probability of the word, obtained through language modeling (a science in itself; suffice it to say that Pr(mushroom soup) > Pr(much rooms hope)); Pr(acoustics word) is the most involved term on the right hand side of the equation and is obtained from the aforementioned hidden Markov models.

TECHNICAL PROBLEMS

Co-articulation of phonemes and words, depending on the input language, can make the task of speech recognition considerably more difficult. In some languages, like English, co-articulator effects are extensive and far-reaching. Consider for example the sentence "what are you going to do?", which when spoken might sound like "whatchagonnado?", which has a phonetic signal which is very different from the expected phonetic signal of each word separately.

ISSN: 2349-7637 (Online)

Intonation and sentence stress can play an important role in the interpretation of an utterance. As a simple example, utterances that might be transcribed as "go!", "go?" and "go." can clearly be recognized by a human, but determining which intonation corresponds to which punctuation is difficult for a computer. Most speech recognition systems are unable to provide any more information about an utterance other than what words were pronounced, so information about stress and intonation cannot be used by the application using the recognizer.

In naturally spoken language, there are no pauses between words, so it is difficult for a computer to decide where word boundaries lie. Some sets of utterances can sound the same, but can only be disambiguated by an appeal to context. A general solution of many of the above problems effectively requires human knowledge and experience, and would thus require advanced pattern recognition and artificial intelligence technologies to be implemented on a computer. In particular, statistical language models are often employed for disambiguation and improvement of the recognition accuracies.

SPEECH IN EDUCATION

Speech-enabled applications and hardware are increasingly finding their way into the classroom and into the offices of educators at all levels of education, but educational applications still represent a small, though growing, and segment of the speech technology market, according to industry analysts.

AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is the process by which a computer maps an acoustic speech signal to text. Automatic speech understanding is the process by which a computer maps an acoustic speech signal to some form of abstract meaning of the speech. Speech synthesis is the task of transforming written input to spoken output. The input can either be provided in a graphemes/orthographic or a phonemic script, depending on its source. As a consequence of its reliance on phonology, linguistics, signal processing, statistics, computer science, acoustics, connectionist networks, psychology and other fields, there are many technologies involved in speech technology.

SPEECH RECOGNITION CUTS PRINTING, PERSONNEL COSTS.

Before installing IntelliSPEECH®, operators had time to do little else than transfer calls. Now callers can simply ask for the faculty member, department or student they want for automatic connection rather than asking the operator to be connected. The university opted for a speech-enabled system rather than a simple touchtone system to replace print directories because it's much simpler and quicker for people to say a name rather than search for the letters of a person's name on a telephone keypad.

In addition to replacing the print directories, the speech-enabled directory enables universities to employ fewer operators and to spend less time training the operators they do have. Additionally, Yellow Pages' ads and university-printed listings can only be changed as often as the directories go to print, and may become outdated shortly thereafter. The speech-enabled auto attendant, on the other hand, enables authorized users to make changes on the fly.

LIMITATIONS OF SPEECH RECOGNITION

Man has been constantly trying to create machines that resemble him. Perhaps, it is the inner urge in man to replicate him that prompts him to make machines that are very similar to him. Speech recognition software has been a major breakthrough that has given us the ability to "talk" with computers. Speech recognition is a very powerful tool indeed, considering the fact that speech is one of the main characteristics that separates humans from animals.

If speakers need to specifically say the punctuations into their computer's mic, that would create a very difficult problem for many, since we are not used to speaking punctuations in real life. This in turn means that the usability and ease of the software is greatly reduced, which can be considered as negative aspects of the software.

The speech recognition software is not at all user friendly even though you only need to take your mic and speak words into it. This is because the speech recognition software cannot distinguish between background sounds and the sound of the user. The software may catch the voice of others speaking in the room, or may catch the voice of computer keystrokes, the sound of other electrical goods etc and type words on the screen that has no relation to what the user might be trying to speak. This makes the software exclusive because it can be used only by CEOs or other officials who may be given a separate room and facilities in the office.

The software has to be spoken to very slowly. Often people who talk fast will see gibberish appearing on their screen because the software cannot decipher the words spoken in such haste. Often speaking slowly will make people lose the tempo with which

ISSN: 2349-7637 (Online)

ideas flow into their minds. Hence, speaking slowly will kill their flow of ideas and may significantly affect their ability to construct proper sentences and coherent ideas.

Speech recognition software cannot measure the emotions with which words are spoken into it. For example, an operator may have to ask the software to underline or capitalize words for effect. This may not happen as easily as said because the emotional situation of the person may not allow him to dictate each and every word.

The software cannot be used by a person who is temperamental since it needs great patience to review each word that is written on the screen because the chances for errors are very high.

A very important drawback with voice recognition packages is that the software has to be trained for each user. This means that in an office where an employee has to work in different machines at different times, training each machine for voice input would be a very costly affair in terms of wasted man hours for doing repetitive work. In addition, during instances of virus attacks or other similar conditions when the hard disk of the machine has to be formatted, the software has to be trained on all machines.

The "trained" software cannot be used for different people since their personal style of speech may be very different. For example, consider a scenario where an employee has left an organization. A new employee cannot use the speech recognition software with as much efficiency with which the previous employees used it because the software has to be trained to suit the voice of the new employee. This means that the office systems and all the associated processes of the office become dependent on one employee and his skills.

It may be seen that issuing voice command to make one's machine do specific tasks does not run into so much trouble as dictating to one's computer to make it write something on screen. This is because the number of words that are used in predefined commands in most software is limited.

FUTURE OF AI FOR SPEECH RECOGNITION

Information and communication technologies are rapidly converging to create machines that understand us, do what we tell them to, and even anticipate our needs. We tend to think of intelligent systems as a distant possibility, but two relentless super trends are moving this scenario toward near-term reality. Scientific advances are making it possible for people to talk to smart computers, while more enterprises are exploiting the commercial potential of the Internet.

Forecasts conducted under the TechCast Project at George Washington University indicate that 20 commercial aspects of Internet use should reach 30% 'take-off' adoption levels during the second half of this decade to rejuvenate the economy.

Meanwhile, the project's technology scanning finds that advances in speech recognition, artificial intelligence, powerful computers, virtual environments, and flat wall monitors are producing a 'conversational' human-machine interface. These powerful trends will drive the next generation of information technology into the mainstream by about 2010.

The following are a few of the advances in speech recognition, artificial intelligence, powerful chips, virtual environments, and flat-screen wall monitors that are likely to produce this intelligent interface. IBM has a Super Human Speech Recognition Program to greatly improve accuracy, and in the next decade Microsoft's program is expected to reduce the error rate of speech recognition, matching human capabilities. MIT (Massachusetts Institute of Technology, US) is planning to demonstrate their Project Oxygen, which features a voice-machine interface. Amtrak, Wells Fargo, Land's End, and many other organizations are replacing keypad-menu call centers with speech-recognition systems because they improve customer service and recover investment in a year or two.

REFERENCES

1. http://research.microsoft.com/en-us/news/features/speechrecognition-082911.aspx
2. http://dl.acm.org/citation.cfm?id=1752355
3. http://www.creativecow.net/interstitial.php?url=http%3A%2F%2Fforums.creativecow.net%2Fthread%2F279%2F626&id=0
4. www.ijsce.org/attachments/File/v2i5/E1054102512.pdf
5. http://en.wikipedia.org/wiki/Outline_of_artificial_intelligence
6. http://www.csd.cs.cmu.edu/research/areas/vis_speech_lang/

ISSN: 2349-7637 (Online)