ACHIEVING SECURITY VIA SPEECH RECOGNITION

Swati Patel and Nita Lokwani

Smt. Chandaben Mohanbhai Patel Institute of Computer Applications,
Charusat University, Changa

ABSTRACT

Speech is one of the essential sources of the conversation between human beings. We as humans speak and listen to each other in human-human interface. People have tried to develop systems that can listen and prepare a speech as persons do so naturally. This paper presents a brief survey on Speech recognition, allow people to compose documents and control their computers with their voice. In other words, the process of enabling a machine (like a computer) to identify and respond to the sounds produced in human speech. ASR can be treated as the independent, computer-driven script of spoken language into readable text in real time. The Speech Recognition system requires careful attention to the following issues: Meaning of various types of speeches, speech representation, feature extraction techniques, speech classifiers, and database and performance evaluation. This paper helps in understanding the technique along with their pros and cons. A comparative study of different technique is done as per stages.

KEYWORDS

Automatic speech recognition (ASR), Analysis, feature extraction, Modeling, Testing, speech processing, Human Computer Interaction (HCI).

1. INTRODUCTION

Speech recognition is a process in which computer (or other type of machine) identifies spoken words and performing required task. The task is to understand spoken words, react appropriately and then convert into text. Sometime it is also known as speech to text (STT). Basically, it is correct recognize of what you are saying. It is mostly used when someone’s hands and eyes are busy. The speech recognition system consists a microphone in which person can speak; speech recognition software to interpret the speech; a good quality sound card for input and/or output; and a proper pronunciation.

2. SPEECH RECOGNITION BASICS

Speech recognition automatically extracts the string of words spoken from the speech signal.

The following factors are the basically used for understanding speech recognition technology.

Utterance

An utterance is the pronunciation (speaking) of either a word or words that represent a single meaning to the computer. It can be word, words, sentence, or even multiple sentences.

DOI : 10.5121/ijist.2016.6222

Speaker dependent & speaker independent systems

Speaker dependent systems design speech patterns for a specific speaker. Generally they are more perfect for the correct speaker, but inaccurate for other speakers. They accept the speaker will speak in a steady and consistent voice and tempo. Speaker independent systems are constructed from a wide range of speakers. Adaptive systems usually start with speaker independent systems and apply training techniques to adapt to the speaker to increase their recognition accuracy.

Vocabularies

It is a collection of words or statements that can be recognized by the SR system. Generally, for a computer, it is very easy to identify smaller vocabularies, but very difficult to identify larger vocabularies. In normal dictionaries, it doesn't have only a single word, but it can be as long as a sentence or two. Smaller vocabularies can remember only one or two words like “Stand up", but large vocabularies can have a hundred thousand or more!

Accuract

The skill of a recognizer can be studied by measuring its accuracy - or how well it remembers the words. This includes not only correctly identifying a word, but also identifying if the spoken a word is not in its vocabulary. Good ASR systems have more than 98% correct. The satisfactory accuracy of a system really based on the application.

Training

Some speech recognizers have the skill to adapt to a speaker. When the system has this skill, it may allow training to take place. An ASR system is trained by having the person who has repeat standard or common phrases and adjusting its comparison algorithms to match that particular person. The accuracy of a recognizer usually improves with training.

Persons who have difficulty in speaking, or pronouncing certain of the words they can also use training. As long as the person can consistently repeat a word, ASR systems with training should be able to adapt.

3. TYPES OF SPEECH RECOGNITION

Speech recognition systems can be divided in various different classes by describing what types of words they have the ability to remember. ASR is the ability to check when a person starts and completes a word. Most packages can fit into more than one class, based on which mode they're using.

Isolated Words

Isolated word required each word to have silent (lack of an audio signal) on both ends. It doesn't mean that it accepts single words, but it does require a single word at a time. Often, these systems have "Listen/Not-Listen" states, where they require the person to wait between words (usually doing processing during the pauses). The isolated word might be a good name for this class.

Connected Words

Connected words (or more correctly 'connected utterances') are similar to isolated words, but allow separate words to be run simultaneously with a minimum pause between them.

Continuous Speech

Processing with continuous speech are some of the most difficult to design because they must process with the special methods to determine word limits. Continuous speech allows users to speak almost naturally, while the computer defines the content. Basically, that particular content itself the words which are dictated by Computer.

Spontaneous Speech

There appears to be a various definitions for what spontaneous speech in reality is. At a basic point, it can be thought of as speech that is natural sounding and not planned. An ASR system with spontaneous speech ability to cover types of natural speech comprises with the words being run together.

Voice Verification/Identification

Some ASR systems are used for identifying specific users. This document doesn't cover the particular verified or secure systems.

4. WORKING OF SPEECH RECOGNITION SYSTEM

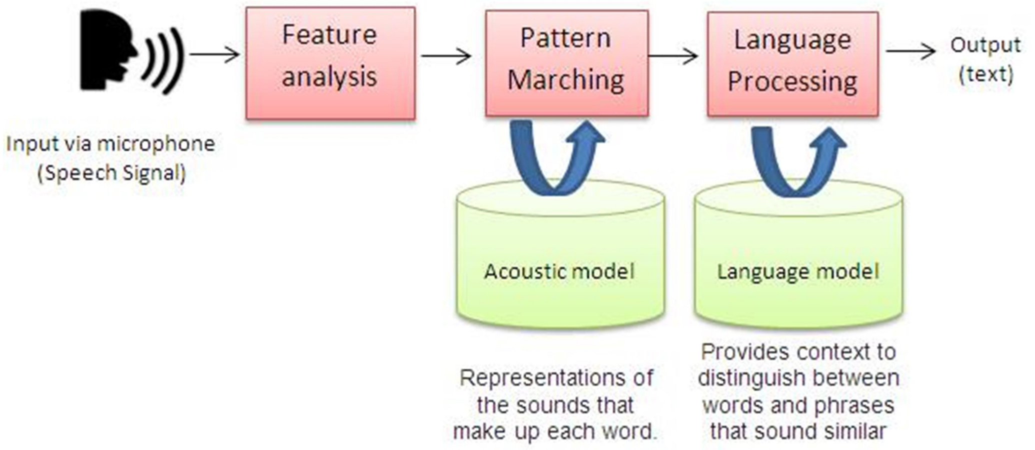
Feature analysis
Pattern Marching
Language Processing
Output (text)
Input via microphone (Speech Signal)
Acoustic model
Language model
Representations of the sounds that make up each word.
Provides context to distinguish between words and phrases that sound similar

Figure 1. Process of Speech Recognition

Basically, the conversion of the voice to an analog signal of microphone and that process takes the input from digital stage. Input of the system from the user is known as utterance (Spoken input from the user of a speech system. An utterance may be a one word, a sentence, an entire phrase, or even several sentences.) This is the representation of the binary form of 1s and 0s that make up programming languages used by the computer. Any other kind of sound is not heard by the computers.

Sound-recognition system has acoustic models (An acoustic model is created by taking audio demos of speech, and their text records, and using software to create numerical representations of the sounds that make up each utterance. It is used by a speech recognition engine to remember speech) convert the audio sounds to one of about four dozen basic speech components (called phonemes). The latest versions of speech technology have been derived so that they eliminate the extra noise and not used information that is not needed to let the computer work. The words we speak are changed into digital forms of the basic speech components (phonemes).

Once this is complete, a second step of the software begins to work. The text is compared to the digital dictionary that is stored in computer internal storage. This is a very vast collection of words, usually more than 100,000. When it compares and identify a match based on the digital form it displays the words on the display. This is the simple and basic process for all speech recognition systems.

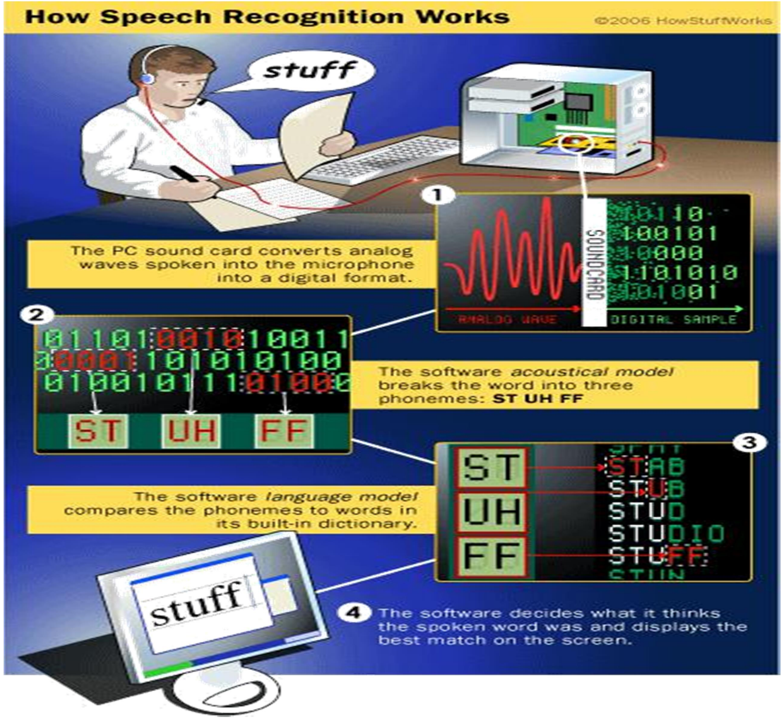
How Speech Recognition Works
©2006 HowStuffWorks
stuff
1
The PC sound card converts analog waves spoken into the microphone into a digital format.
SOUNDCARD
ANALOG WAVE
DIGITAL SAMPLE
2
ST
UH
FF
The software *acoustical model* breaks the word into three phonemes: **ST UH FF**
3
STAB
STUB
STUD
STUDIO
STUFF
The software *language model* compares the phonemes to words in its built-in dictionary.
stuff
4
The software decides what it thinks the spoken word was and displays the best match on the screen.

Figure 2. Working of Speech Recognition

To convert speech to text or a computer command, a computer has to go with the several complex tasks. When you speak, you create vibrations in the air. The analog-to-digital converter (ADC) converts this analog wave to the digital data that can be understood by computer. To do this digitizes the sound by taking small measurements of the wave at the regular time. The system filters the digitized sound to remove unwanted noise, and sometimes to separate it into various bands of frequency (frequency is the wavelength of the sound waves, heard by humans as differences in pitch). It also normalizes the sound, or adjusts it to a constant volume level. It may also have to be temporally aligned. The speed of the speech has been always in variable form, so the sound must be adjusted to match the speed of the format sound that is already available in the system's memory.

Next the signal is divided into small segments as short as a few hundreds of a second, or even thousandths in the case of passive consonant sounds -- consonant stops produced by obstructing airflow in the vocal tract -- like "p" or "t." The program then matches these segments to known phonemes in the appropriate language. A phoneme is the smallest element of a language -- a representation of the sounds we make and put together to form meaningful expressions. There are roughly 40 phonemes in the English language (different linguists have different opinions on the exact number), while other languages have more or fewer phonemes.

The next step seems simple, but it is actually the most difficult to accomplish and is the focus of most speech recognition research. The program examines phonemes in the context of the other phonemes around them. It runs the contextual phoneme plot through a complex statistical model and compares them to a large library of known words, phrases and sentences. The program then determines what the user was probably saying and either outputs it as text or issues a computer command.

5. APPROACHES TO SPEECH RECOGNITION

- Acoustic phonetic approach
- Pattern Recognition approach
- Artificial intelligence approach.

5.1 Acoustic Phonetic Approach

It is also called as rule-based approach. It is use knowledge of phonetics & linguistics to guide the search process. This approach generally uses some principles which are defined expressing everything or anything that might help to decrypt based in “blackboard” architecture, i.e. At each decision point it lays out the possibilities and use rules to determine which orders are acceptable. It has poor performance due to difficulty to express rules, to improve the organization. This approach recognizes individual phonemes, words, sentence structure and/or significance.

5.2 Pattern Recognition Approach

This method is divided in two steps, i.e. training of speech patterns and recognition of pattern by way of pattern comparison. In the parameter measurement phase (filter bank, LFC, DFT), a sequence of measurements is developed based on the input signal to define the “test pattern”.

The unknown test pattern is then compared with each sound reference pattern and a measure of comprise between the examination pattern & reference pattern. Best matches the unknown test pattern based on the matching of the pattern classification phase (dynamic time warping).

5.2.1 Template-based approach

This approach provides a family of techniques that have advanced the field considerably during the last six decades. A collection of prototypical speech patterns is stored as reference patterns representing the dictionary of applicant’s words. Recognition is then carried out by matching an unknown spoken utterance with each reference template and choosing the category of the best matching pattern. Generally templates are created for entire words. This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided.

5.2.2 Statistics-based approach

Stochastic modelling entails the use of probabilistic models to deal with undefined or incomplete information. In speech recognition, many sources like, confusable sounds, speaker variability’s, contextual effects, and homophone words are affected for uncertainty and incompleteness. In today the most popular stochastic approach is hidden Markov modelling. A hidden Markov model is qualified by a finite state Markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variabilities, while the parameters in the output distribution model, spectral variabilities. These two types of variables are the effect of speech recognition.

5.3 Artificial Intelligence Recognition Approach

This approach is a combination of acoustic phonetic approach and pattern recognition approach. In this, it exploits the concepts of acoustic phonetics and pattern recognition methods. The information regarding linguistic, phonetic and spectrogram used by Knowledge based approach. Some speech researchers developed recognition system that used acoustic phonetic knowledge to improve classification rules for speech sounds. While template based approaches have been real in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby creating error analysis and knowledge-based system enhancement difficult. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding of human spoken language processing. In its complete form, the knowledge engineering plan involves the direct and explicit incorporation of expert's speech knowledge into a recognition system. This knowledge is normally derived from careful study of spectrograms and is incorporated using guidelines or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems.

6. CHALLENGES AND DIFFICULTIES OF SR

Speech Recognition is still a very cumbersome problem. Following are the problems….

- **Speaker Variability**

  Speaker or more than one speaker may be pronounced the same word in a different way.

- **Channel Variability**

  The position and quality of the microphone and background environment may be affecting in output

7. APPLICATIONS OF SPEECH RECOGNITION

Applications of Speech Recognition

Speech recognition applications include

- Voice dialling (e.g., "Call home"),
- Call routing (e.g., "I would like to make a collect call"),
- Simple data entry (e.g., entering a credit card number),
- Preparation of structured documents (e.g., A radiology report),
- Speech-to-text processing (e.g., word processors or emails), and
- In aircraft cockpits (usually termed Direct Voice Input).

8. PROS AND CONS OF SPEECH RECOGNITION

PROS:

- There is no need to type or write text, and generally it is quicker than “typing” & “handwriting”.

- Allows for better spelling, whether it is in text or documents. It is very useful for mental or physical disability person.

CONS:

- No program is 100% perfect

- Factors like slang, homonyms, signal-to-noise ratio, and overlapping speech are affecting the accuracy of speech recognition.

- Can be expensive depending on the program

9. CONCLUSIONS

Speech is the primary, and the most appropriate means of communication between human being. Whether due to technological curiosity to make machines that mimic humans or desire to automate work with machines, research in speech and speaker identification. This paper introduces the basics of speech recognition technology and also highlights the difference between different speech recognition systems. In this paper the most common algorithms which are used to do speech recognition are also discussed along with the current and its future use. Speech recognition is one of the most integrating areas of machine intelligence, since, humans do a daily activity of speech recognition.

10. ACKNOWLEDGEMENT

We are very thankful to our institute CMPICA (Smt. Chandaben Mohanbhai Patel Institute of Computer Applications), CHARUSAT for providing necessary infrastructure for development of the research work.

REFERENCES

[1] M.A.Anusuya, S.K.Katti "Speech Recognition by Machine: A Review" International Journal of Computer Science and Information Security

[2] Preeti Saini, Parneet Kaur "Automatic Speech Recognition: A Review" International Journal of Engineering Trends and Technology.

[3] Santosh k.Gaikwad, Bharti W.Gawali, Pravin Yannawar "A Review on Speech Recognition Technique" International Journal of Computer Applications.

[4] Parwinder pal Singh, Er. Bhupinder Singh "Speech Recognition as Emerging Revolutionary Technology, "International Journal of Advanced Research in Computer Science and Software Engineering.

[5] http://www.tldp.org/HOWTO/Speech-Recognition-HOWTO/introduction.html

[6] Shipra J. Arora, Rishi Pal Singh "Automatic Speech Recognition: A Review", International Journal of Computer Applications.

[7] Wiqas Ghai and Navdeep Singh,"Literature Review on Automatic Speech Recognition",International Journal of Computer Applications vol. 41– no.8, pp. 42-50, March 2012.

[8] Nnamdi Okomba S., 2Adegboye Mutiu Adesina, and 3Candidus O. Okwor., "Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research", IOSR Journal of Electronics and Communication Engineering

[9] http://www.computerhope.com/jargon/v/voicreco.htm

[10] https://en.wikipedia.org/wiki/Speech_recognition

[11] https://en.wikipedia.org/wiki/Acoustic_model

AUTHORS

Swati Patel received her M.C.A. & B.C.A. degrees from Dharmsinh Desai University, Nadiad, Gujrat, India in 2011 & 2009 respectively. She is presently working as an Assistant Professor in Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Changa. Her research area includes HCI and Pattern Recognition.

Nita Lokwani received her M.C.A. degree from Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charusat University, Changa, Gujrat, India in 2012 & B.C.A. degree from Dharmsinh Desai University, Nadiad, Gujrat, India in 2009. She is presently working as an Assistant Professor in Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Changa. Her research area includes Human Computer Interaction and working on Embedded Systems.