Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 88 (2015) 36505-36509

ISSN: 2229-712X

Speech recognition - a computer mediated approach

Sreedhar Appalabatla[1], Mungamuru Nirmala[2] and Kaliyaperumal Karthikeyan[2]

[1]Department of Electrical & Computer Engineering, Hawassa University, Hawassa, Ethiopia, N.E.Africa

[2]Department of Computer Science, Eritrea Institute of Technology, Asmara, Eritrea, N.E.Africa

ARTICLE INFO

Article history:

Received: 25 June 2012;
Received in revised form:
15 November 2015;
Accepted: 20 November 2015;

Keywords

Artificial Intelligence,
Speech Technology,
Strategic Modeling,
Neural network,
Traditional Linguistics,
Automatic Speech Recognition (ASR),
Hidden Markov Models (HMM).

ABSTRACT

The computer revolution is now well advanced, but although we see a starting preparation of computer machines in many forms of work people do, the domain of computers is still significantly small because of the specialized training needed to use them and the lack of intelligence in computer systems. In the history of computer science five generations have passed by, each adding a new innovative technology that brought computers nearer and nearer to the people. Now it is sixth generation, whose prime objective is to make computers more intelligent i.e., to make computer systems that can think as humans. The fifth generation was aimed at using conventional symbolic Artificial Intelligence techniques to achieve machine intelligence. But it failed. Statistical modeling and Neural Networks falls in really sixth generation. The goal of work in Artificial Intelligence is to build the machines that perform tasks normally requiring human intelligence. Its true, but speech recognition, seeing and walking doesn't require "intelligence", but human perceptual ability and motor control.

Introduction

Speech Technology is now one of the major significant scientific research fields under the broad domain of AI; indeed it is a major co-domain of computer science, apart from the traditional linguistics and other disciplines that study the spoken languages. The days when you had to keep starring at the computer screen and frantically hit the key or click the mouse for the computer to respond to your commands may soon be the things of past. Today one can stretch out and relax and tell their computer to do bidding. This has been made possible by the ASR (Automatic Speech Recognition**)** technology.

The ASR technology would be particularly welcomed by automated telephone exchange operators, doctors and lawyers, besides others whose seek freedom from tiresome conventional computer operations using keyboard and the mouse. It is suitable for applications in which computers are used to provide routine information and services. The ASR's direct speech to text dictation offers a significant advantage over traditional transcriptions. With further refinement of the technology, text will become a thing of past. ASR offers a solution to this fatigue-causing procedure by converting speech in to text.

This paper discusses the concept of Speech Recognition, how the ASR Technology works, followed by classification of speech recognition systems. An attempt is also made to identify the variations in speech, vocabularies for computers, advantages of speaker independent system and speaker dependent system.

Speech Technology is now one of the major significant scientific research fields under the broad domain of AI; indeed it is a major co-domain of computer science, apart from the traditional linguistics and other disciplines that study the spoken languages.

The ASR technology is presently capable of achieving recognition accuracies of 95% - 98 % but only under ideal conditions. The technology is still far from perfect in this uncontrolled real world. The routes of this technology can be traced to 1968 when the term Information Technology hadn't even been coined. American's had only begun to realize the vast potential of computers. In the Hollywood blockbuster 2001: a space odyssey, a talking listening computer HAL-9000, had been featured, a figure in both science fiction and in the world of computing. Even today almost every speech recognition technologist dreams of designing an HAL-like computer with a clear voice and the ability to understand normal speech. Though the ASR technology is still not as versatile as HAL, it can nevertheless be used to make life easier. New application specific standard products, interactive error-recovery techniques, and better voice activated user interfaces allow the handicapped, computer-illiterate, and rotary dial phone owners to talk to the computers. ASR, by offering a natural human interface to computers, finds applications in telephone-call centers, such as for airline flight information system, learning devices, toys, etc.

How does the ASR Technology work?

When a person speaks, compressed air from the lungs is forced through the vocal tract as a sound wave that varies as per the variations in the lung pressure and the vocal tract. This acoustic wave is interpreted as speech when it falls upon a person's ear. In any machine that records or transmits human voice, the sound wave is converted into an electrical analogue signal using a microphone. When we speak into a telephone receiver, for instance, its microphone converts the acoustic wave into an electrical analogue signal that is transmitted through the telephone network. The electrical signals strength of microphone varies in amplitude over time and is referred to as an analogue signal or an analogue waveform. If the signal results from speech, it is known as a speech waveform. Speech waveforms have the characteristic of being continuous in both time and amplitude.

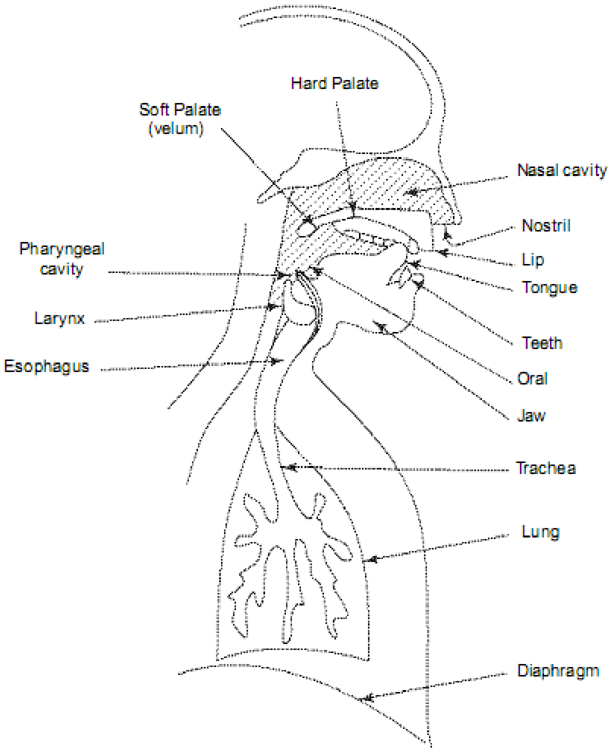
Tele:
E-mail addresses: appalabatla.s@gmail.com

Listener's ears and brain receive and process the analogue speech waveforms to figure out the speech. ASR enabled computers, too, work under the same principle by picking up acoustic cues for speech analysis and synthesis. Because it helps to understand the ASR technology better, let us do well a little more on the acoustic process of the human articulator system. In the vocal tract the process begins from the lungs. The variations in air pressure cause vibrations in the folds of skin that constitute the vocal chords. The elongated orifice between the vocal chords is called the glottis. As a result of the vibrations, repeated bursts of compressed air are released into the air as sound waves.

Articulators in the vocal tract are manipulated by the speaker to produce various effects. The vocal chords can be stiffened or relaxed to modify the rate of vibration, or they can be turned off and the vibration is eliminated while still allowing air to pass. The velum acts as a gate between the oral and the nasal cavities. It can be closed to isolate or opened to couple the two cavities. The tongue, jaw, teeth, and lips can be moved to change the shape of the oral cavity.

The nature of sound preserves the wave radiating out to the world from the lips depends upon time varying articulations and the absorptive qualities of the vocal tracts materials. The sound pressure wave exists as a continually moving disturbance of air. Particles come move closer together as the pressure increases or move further apart as it decreases, each influencing its neighbor in turn as the wave propagates at the speed of sound. The amplitude to the wave at any position, distant from the speaker, is measured by the density of air molecules and grows weaker as the distance increases. When this wave falls upon the ear it is interpreted as sound with discernible timbre, pitch, and loudness.

Fig 1. Speech Production Physiology



Air under pressure from the lung moves through the vocal tract and comes into contact with various obstructions including palate, tongue, teeth, lips and timings. Some of its energy is absorbed by these obstructions; most is reflected. Reflections occur in all directions so that parts of waves bounce around inside the cavities for some time, blending with other waves, dissipating energy and finally finding the way out through the nostrils or past the lips. Some waves resonate inside the tract according to their frequency and the cavity's shape at that moment, combining with other reflections, reinforcing the wave energy before exiting. Energy in waves of other, non-resonant frequencies are attenuated rather than amplified in its passage through the tract.

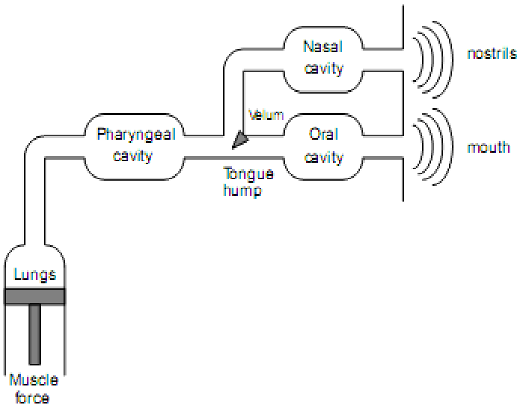
You speak into a microphone, which converts sound waves into electrical signals. The ASR program removes all noise and retains only the words that you have spoken. The words are broken down into individual sounds, known as phonemes, which are the smallest sound units discernible.

In the next most complex part, the program's database maps sounds to character groups. The program also has a big dictionary of popular words that exist in the language. Next, each phoneme is matched against the sounds and converted into the appropriate character group. This is where problems begin. To overcome the difficulties encountered in this phase, the program uses numerous methods; firs, it checks and compares words that are similar in sound with what they have heard; then it follows a system of language analyses to check if the language allows a particular syllable to appear after another.

Then comes the grammar and language check. It tries to find out whether or not the combination of word makes any senses. This is very similar to the grammar-check package that you find in word processors.

The numerous words constituting the speech are finally noted down in the word processor, while all speech recognition programs come with their own word processors, some can work with other word processing packages like MS word and Word Perfect. In fact, OS/2 allows even operating command to be spoken.

Fig 2. A Block Diagram of Human Speech Production



The Speech Recognition Process

When a person speaks, compressed air from the lungs is forced through the vocal tract as a sound wave that varies as per the variations in the lung pressure and the vocal tract. This acoustic wave is interpreted as speech when it falls up on a person's ear. Speech waveforms have the characteristic of being continuous in both time and amplitude.

Any speech recognition system involves five major steps:

i Converting sounds into electrical signals: when we speak into microphone it converts sound waves into electrical signals. In any machine that records or transmits human voice, the sound wave is converted into an electrical signal using a microphone. When we speak into telephone receiver, for instance, its microphone converts the acoustic wave into an electrical analogue signal that is transmitted through the telephone network. The electrical signal's strength from the microphone varies in amplitude overtime and is referred to as an analogue signal or an analogue waveform.

ii Background noise removal: The ASR program removes all noise and retains the words that you have spoken.

iii Breaking up words into phonemes: The words are broken down into individual sounds, known as phonemes, which are the smallest sound units discernible. For each small amount of time, some feature, value is found out in the wave. Likewise, the wave is divided into small parts, called Phonemes.

iv Matching and choosing character combination: This is the most complex phase. The program has big dictionary of popular words that exist in the language. Each Phoneme is matched against the sounds and converted into appropriate character group. This is where problem begins. It checks and compares words that are similar in sound with what they have heard. All these similar words are collected.

v Language analysis: Here it checks if the language allows a particular syllable to appear after another.

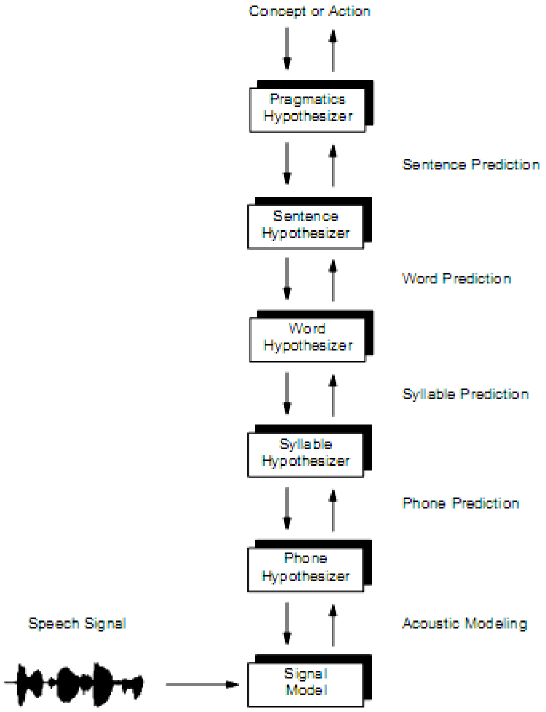
Variations in Speech

The speech-recognition process is complicated because the production of phonemes and the transition between them varies from person to person and even the same person. Different people speak differently. Accents, regional dialects, sex, age, speech impediments, emotional state, and other factors cause people to pronounce the same word in different ways. Phonemes are added, omitted, or substituted. The rate of speech also varies from person the person depending upon a person's habit and his regional background.

A word or a phrase spoken by the same individual differs from moment to moment. Illness, tiredness, stress or other conditions cause subtle variations in the way a word is spoken at different times.

Also, the voice quality varies in accordance with the position of the person relative to the microphone, the acoustic nature of the surroundings, or the quality of the recording devices. The resulting changes in the waveform can drastically affect the performance of the recognizer.

Fig 3. A Generic Solution



Classification of speech recognition system

When a speech recognition system requires a word to be spoken individually in isolation from other words, it is said to be an isolated word system, it recognizes discrete words only. When they are separated from their neighbors by distinct inter word gap. So the user has to pause before uttering the next word. A continuous word system on the other hand is capable of handling more fluent speeches, while a big vocabulary system works with a vocabulary base with minimum thousand words and a small vocabulary system works with a maximum thousand words.

Vocabularies for Computers

Each ASR system has LAN active vocabulary- a set of words from which the recognition engine tries to make sense of utterance- and a total vocabulary size - the total number of words in all possible sets that can be called from the memory. The vocabulary size and system recognition latency - the allowable time to accurately recognize an utterance determines the process horsepower of the recognition engine.

An active vocabulary set comprises approximately fourteen words plus none of the above, from which the recognizer chooses when none of the fourteen words is good match .The recognition latency when using a 4-MIPS processor, is about.5 seconds for independent set. Processing power requirements increased dramatically for LVR sets with thousands of words. Real time latencies with a vocabulary base of few thousands are possible only through the use of Pentium class processors. A small active vocabulary limits a system search range providing advantages in latency and search time. A large total vocabulary enables more versatile human interface but affects system memory requirements. A system with a small active vocabulary with each prompt usually provides faster more accurate results, similar sounding words in vocabulary set cause recognition errors. But a unique sound for each word enhances recognition engines accuracy.

Which System to Choose

In choosing speech recognition system you should consider the degree of speaker independence it offers. Speaker independent systems can provide high recognition accuracies for a wide range of users without a need to adapt to each user's voice. A pre-requisite of Speaker dependent systems is to train the system with your voice to attain high accuracy. Speaker adaptive systems - an intermediate category are essentially speaker-independent but can adapt their templates for each user to improve accuracy.

Advantages of Speaker Independent System

The advantage of a speaker independent system is obvious. Anyone can use the system without first training it. However, its drawbacks are not so obvious. One limitation is the work that goes into creating the vocabulary templates. To create reliable speaker-independents templates, someone must collect and process numerous speech samples.

This is a time-consuming task; creating these templates is not a one-time effort. Speaker-independent templates are language-dependant, and the templates are sensitive not only to two dissimilar languages but also to the differences between British and American English. Therefore, as part of your design activity, you would need to create a set of templates for each language or a major dialect that your customers use. Speaker independent systems also have a relatively fixed vocabulary because of the difficulty in creating a new template in the field at the user's site.

Advantages of Speaker Dependent System

A speaker dependent system requires the user to train the ASR system by providing examples of one's own speech. Training can be tedious process, but the system has the advantage of using templates that refer only to the specific user and not some vague average voice. The result is language independence. You can say ja, si, or ya during training, as long as you are consistent. The drawback is that the speaker-dependent system must do more than simply matching the incoming speech to the templates. It must also include resources to create those templates.

Comparative Analysis

For a given amount of processing power, a speaker dependent system tends to provide more accurate recognition than a speaker-independent system. A speaker independent system is not necessarily better: the difference in performance stems from the speaker independent template encompassing wide speech variations.

Techniques in vogue

The most frequently used speech recognition technique involves template matching, in which vocabulary words are characterized in memory a template time based sequences of spectral information taken from waveforms obtained during training.

As an alternative to template matching, feature based designs have been used in which a time sequence of the pertinent phonetic features is extracted from a speech waveform. Different modeling approaches are used, but models involving state diagrams have been fond to give encouraging performance. In particular, HMM (Hidden Markov Models) are frequently applied. With HMM's any speech unit can be modeled; all knowledge sources can be modeled, and can be included in a single, integrated model. Various types of HMM's have been implemented with differing results. Some model each word in the vocabulary, while others' model the sub word speech units also.

Hidden Markov Model

HMM can be used to model an unknown process that produces a sequence of observable outputs at discrete intervals, where the outputs are members of some finite alphabet. These models are called Hidden Markov models precisely because the state sequence that produced the observable output is not known - it's hidden.

HMM is represented by a set of states, vectors defining transitions between certain pairs of those states, probabilities that apply to state to state transitions, sets of probabilities characterizing observed output symbols, and initial conditions.

What Lies Ahead?

If we were to draw a path that tell us where speech and computers go from here on, we would see that we have only moved two steps.

Earlier systems would be able to listen to sounds and convert them into words. That was the era of discrete speech programs that listen to each distinct word spoken by you and then put them down. Though slow and tedious, it announced the arrival of speech technologies on the desktop.

Subsequently continuous speech recognition concept came into picture. One can no longer have to speak out each word separately instead they can just talk naturally and the program understands what has been said

The third step is speech understanding. This technology is the one that will actually mark the biggest change in the way we use our computers. It will necessitate a complete overhaul of our operation system, word processors, and spreadsheets just about everything. It will also mark the emergence of a computer like HAL. When speech understanding arrives in its true form, it will allow your computer to make sense of commands like 'wake me up at six in the morning, you ghonchu.' Till then we must only wait and watch.

Need for Building an Ideal System

Speech recognition systems would ideally have to recognize continuous words in a given language, spoken by any person, 100 percent of the time. This is better than most humans can do, especially given the variety of different accents which people have. The accuracy of speech recognition systems grows and matures with time.

The best speech recognition systems today generally either can recognize small vocabularies over a diverse group of speakers, or large vocabularies for only one speaker. A speaker dependent system can be custom built for a single speaker to recognize isolated words or continuous speech. Clearly, the ideal system would recognize continuous speech over a large vocabulary and be speaker independent. Unfortunately, we have no yet succeeded in building such a system. Currently, speech recognition systems are bet at recognizing long words with many distinct features. The easiest systems to build are speaker dependent and isolated word recognition systems. With a drop in the equipment cost and increase in the processing speed of processors, speech recognition products will soon be playing a major role in more demanding jobs.

Advantages

i Security: With this technology a powerful interface between man and computer is created as the voice reorganization understands only the prerecorded voices and hence there are no ways of tampering data or breaking the codes if created.

ii Productivity: It decreases work as all operations are done through voice recognition and hence paper work decreases to its maximum and the user can feel relaxed irrespective of the work.

iii An aid for handicapped and blind: This technology is great boon for blind and handicapped as they can utilize the voice recognition technology for their works.

iv Increase in usability of other languages: As the speech recognition technology needs only voice and irrespective of the language in which it is delivered it is recorded, due to this perspective this is helpful to be used in any language.

v Creation of Personal voice macros: Every day tasks like sending mails receiving mails drafting documents can be done easily and also many tasks speed can be increased.

Conclusion

Voice recognition promises rosy future and offer wide variety of services. The next generation of voice recognition technology consists of something called neural networks using artificial intelligence technology. They are formed by interconnected nodes which do parallel processing of the input for fast evaluation. Like human beings they learn new pattern of speech automatically.

References

1.J.K. Baker, "The dragon system – An Overview," IEEE Trans. Acoustics, Speech Signal Processing, vol, ASSP-23 no.1, pp.24-29, Feb, 1975.

2. F.Jelinek, "Continuous speech recognition by statistical methods," Proc. IEEE, vol.64, pp.532-536, Apr.1976.

3.R.Bakis, "Continuous speech word recognition via centi - second acoustic states," in Proc. ASA Meeting (Washington, DC), Apr.1976.

4.A.B.Poritz, "Linear Predicative hidden Markov models and the speech signal, "in Proc.ICASSP'82 (Paris, France), pp.1291-1294. May 1982.

5.L.R.Rabiner, B.H.Juang. S.E.Levinson, and M.M.Sondhi, "Some properties of continuous hidden Markov model representations, "AT&T Tech.J., vol 64, no 6, pp. 1251-1270, July Aug.1985.

6.R.Schwartz et al., "Context-dependent modeling for acoustic phonetic recognition of continuous speech, "in Conf. Proc. IEEE int. Conf. on Acoustics, Speech and signal Processing, pp.1205-1208, Apr.1985.

7.Rabiner, Lawrence and Juang, Biing-Hwang; Fundamentals of Speech recognition;

8. Lea, Wayne A.; Trends in Speech Recognition

9.Russell and Norvig; Artificial Intelligence, A Modern Approach; 1995.

10.Saito, Shuzo and Nakata, Kazuo fundamentals of Speech Signal Processing; 1985

11.D.B.Paul, "A speaker stress resistant HMM isolated word recognizer, "in Proc. ICASSP'87(Dallas, TX), pp.713-716, Apr 1987.

12.A.M.Derouault, "Context dependent phonetic Markov models for large vocabulary speech recognition," in Proc ICASSP'87(Dallas, TX), Paper 10.1.1, pp 360-363, Apr.1987.

13.S.Furui, "Speaker independent isolated word recognition based on dynamics emphasized cepstrum," Trans. ICE of Japan, vol.69, no.12. pp. 1310-1317, Dec.1986.