International Journal of Applied Research & Studies ISSN 2278 – 9480

Research Article

Overview of Speech Recognition and Recognizer

Authors

[1]Dr. E. Chandra, [2]Dony Joy

Address for Correspondence:

[1] Director, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore

[2]Research Scholar, D J Academy for Managerial Excellence, Coimbatore

Abstract

This paper discuss about the concepts of speech recognition, recognizer and their performance. Speech is a natural mode of communication for people. People learn all the relevant skills during early childhood, without instruction, and they continue to rely on speech communication throughout their lives. Speech recognition is the task of converting any speech signal into its orthographic representation.

Keywords: speech, communication, recognition, recognizer.

1. Introduction

Speech technology and computing power have created a lot of interest in the practical application of speech recognition. Speech is the primary mode of communication among humans [1] [2]. Our ability to communicate with machines and computers, through keyboards, mice and other devices, is an order of magnitude slower and more cumbersome. In order to make this communication more user- friendly, speech input is an essential component [2]. Recognition covers a number of different approaches of creating software which enable computers to recognize natural human speech. Though related in concept to computers that can repeat the spoken words, technology that makes us to speak, but the computer works quite differently [3] [4].

There are broadly three classes of speech recognition applications

First, isolated word recognition systems, each word is spoken with pauses before and after it, so that end-pointing techniques can be used to identify word boundaries reliably [5].

Secondly, highly constrained command-and-control applications use small vocabularies, limited to specific phrases, but use connected word [5] or continuous speech.

Finally, large vocabulary continuous speech systems have vocabularies of several tens of thousands of words, and sentences can be arbitrarily long, spoken in a natural fashion. The last is the most user-friendly but also the most challenging to implement. However, the most accurate speech recognition systems in the research world are still far too slow and expensive to be used in practical, large vocabulary continuous speech applications on a wide scale.

2. Speech recognition – Voice commands

Instead of moving mouse, voice commands can be given to the computer, telling it to open menus and choose commands. Commands can be issued to edit and format by voice, cut and paste, open programs, close windows, create e-mail, and surf the web [6].

3. Speech Recognition Procedure

The speech recognition process [5] is divided into various phases, which is illustrated in Figure 1.1.

.

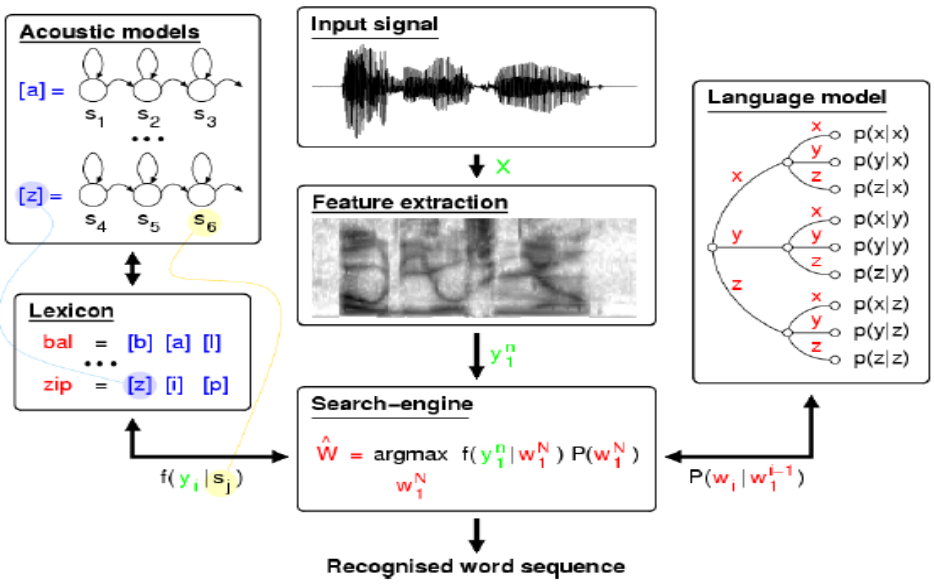


Figure 1.1: Speech Recognition Process

Input Signal

The speaker's voice is captured from an input device and is converted from analog to digital speech signal [5] [7], which form the Input signal. The commonly used input device is a microphone. Note that, the quality of the input device can influence the accuracy of Speech Recognition system. The same applies to acoustic environment. For instance, additive noise, room reverberation, microphone position and type of microphone can all relate to this part of process.

Feature Extraction

The next block, which shows the feature extraction subsystem, is to deal with the problems created in the first part, as well as deriving acoustic representations [4]. The two aims are, separate classes of speech sounds, such as music and speech, and effectively suppress irrelevant sources of variation.

Search Engine

The search engine block is the core part of speech recognition process. In a typical Automatic Speech Recognition (ASR) system, a representation of speech, such as spectral or cepstral representation is computed over successive intervals, for example, 100 times per second. These representations or speech frames are then compared with the spectra frames, which were used for training. It is done by using some measurement of distance of similarity. Each of these comparisons can be regarded as a local match. The global match is a search for the best sequence of words, in the sense that is the best match to the data and it is determined by integrating many local matches.

The local match does not usually produce a single hard choice of the closet speech class, but rather a group of distances or probabilities corresponding to possible sounds. These are then used as part of a global search or

decoding to find an approximation to the closest sequence of speech classes, or ideally to the most likely sequence of words.

Acoustic Model

Speech recognition and language understanding are two major research thrusts that have traditionally been approached as problems in linguistics and acoustic phonetics, where a range of acoustic-phonetic knowledge has been brought to bear on the problem with remarkably little success [8].

One of the key issues in acoustic modeling has been the choice of a good unit of speech. Small vocabulary systems of a few tens of words, it is possible to build separate models for entire words, but this approach quickly becomes infeasible as the vocabulary size grows. For one thing, it is hard to obtain sufficient training data to build all individual world models. It is necessary to represent words in terms of sub-word units, and train acoustic models for the latter, in such a way that the pronunciation of new words can be defined in terms of already trained sub-word units[9].

The phoneme (or phone) has been the most commonly accepted sub-word unit. There are approximately 50 phones in spoken English language; words are defined as sequences of such phones. Each phone is, in turn, modeled by a Hidden Morkov Model (HMM). Natural continuous speech has strong co-articulatory effects. Informally, a phone models the position of various articulators in the mouth and nasal passage (such as the tongue and the lips) in the making of a particular sound.

Since these articulators have to move smoothly between different sounds in producing speech, each phone is influenced by the neighboring ones, especially during the transition from one phone to the next [10]. This is not a major concern in small vocabulary systems in which words are not easily confusable, but becomes an issue as the vocabulary size and the degree of confusability increase.

The acoustic model is the recognition system's model for the pronunciation of words, crucial to translating the sounds of speech to text. In reality, the type of speaker model used by the recognition engine greatly affects the type of acoustic model used by the recognition system to convert the vocalized words to data for the language model (Context / Grammar) to be applied.

There are a wide variety of methods to build the pattern models, the three major types are:

a. Vector Quantization (VQ)
b. Hidden Markov Models (HMM)
c. Neural Networks

5. The Speech Advantage

There are fundamentally three major reasons why so much research and effort has gone into the problem of trying to teach machines to recognize and understand fluent speech, and these are the following:

Cost reduction

Among the earliest goals for speech recognition systems was to replace humans, who were performing some simple tasks, with automated machines, thereby reducing labor expenses while still providing customers with a natural and convenient way to access information and services. One simple example of a cost reduction system was the Voice Recognition Call Processing (VRCP) system introduced by AT&T in 1992 which is essentially automated so-called “Operator Assisted” calls, such as Person-to-Person calls, Reverse billing calls, Third Party Billing calls, Collect Calls (by far the most common class of such calls), and Operator-Assisted Calls.

New revenue opportunities

Speech recognition and understanding systems enabled service providers to have a 24x7 high quality customer care automation capability, without the need for access to information by keyboard or touch tone button pushes. The first example of such a service was, the How May I Help You (HMIHY) service introduced by AT&T late in 1999 which automated the customer care for AT&T Consumer Services. Second example of such a service was the NTT Anser service for voice banking in Japan, which enabled Japanese banking customers to access bank account records from an ordinary telephone without having to go to the bank. Of course, today the users utilize the Internet for such information, but in 1988, when this system was introduced, the only way to access such records was a physical trip to the bank and a wait in lines to speak to a banking clerk.[10][8]

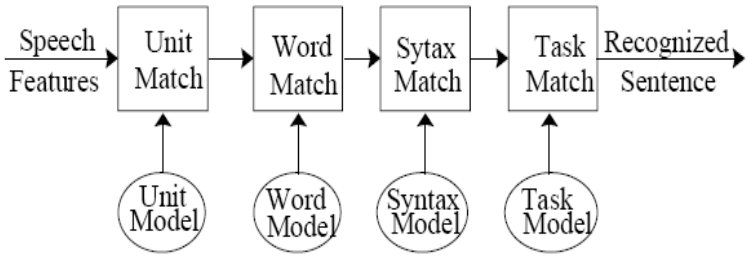
Customer retention

Speech recognition provides the potential for personalized services based on customer preferences, and thereby to improve the customer experience. A trivial example of such a service is the voice-controlled automotive environment which recognizes the identity of the driver from voice commands and adjusts the automobile's features (seat position, radio station, mirror positions, etc.) to suit the customer preference.

6. Why is Speech Recognition Difficult?

Speech is actually, Time-varying Signal, which is well-structured communication process, Depends on known physical movements, Composed of known, distinct units and modified when speaking to improve Signal to Noise ratio.

A typical modular continuous speech recognizer is shown in the below figure.



7. Speech Recognizer

There are number of basic recognizer's are available in the marker which is either commercial or Non-commercial.

- **Commercial Recognizer's are**
  - IBM's ViaVoice (Linux, Windows, MacOS)
  - Dragon Naturally Speaking (Windows)
  - Microsoft's Speech Engine SAPI (Windows)
  - BaBear (Linux, Windows, MacOS)
  - SpeechWorks (Linux, Sparc & x86 Solaris, Tru64, Unixware, Windows)

- **Non-commercial Recognizer's are**

  - OpenMind Speech (Linux)
  - XVoice (Linux)
  - CVoiceControl/kVOiceControl (Linux)
  - GVoice (Linux)
  - Sphinx (Windows,Linux, Mac os)
  - Julius Speech Recognition Engine
  -

8. **Applications of speech recognition**

Some of the applications of speech recognition are

1. Data Entry Enhancements in an Electronic Patient Care Report (ePCR).
2. Dictation.
3. Command and Control.
4. Telephony.
5. Wearable.
6. Medical/Disabilities.
7. Embedded Applications.
8. Agricultural application to get farmer queries.

9. Who can benefit from Speech Recognition?

- Persons with mobility impairments or injuries that prevent keyboard access
- Persons who have or who are seeking to prevent repetitive stress injuries
- Persons with writing difficulties
- Any person who want hands-free access to the computer
- Any persons who wants to increase their typing speed (reportedly up to 150 wpm)[11]

10. Conclusion

In this paper, an overview study and analysis of the concepts of speech recognition and recognizer was discussed. The future work will be the study about the quality of rate of speech and accuracy by new hybrid speech recognizer algorithms.

References

1. Cantor, A. (2001). Speech recognition: An accommodation planning perspective. Proceedings of CSUN 2001 Conference, Los Angeles. Northridge: California State University.
2. Paul. D.B, "An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model," in Proc. ICASSP '92, San Francisco, CA, pp. 25-28, Mar. 1992.
3. Kubala F, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated Technologies for Indexing Spoken Language," Communications of the ACM 43, No. 2, 48-56, February 2000.
4. Michael Witbrock and Alexander G. Hauptmann,"Speech Recognition and Information Retrieval Experiments In Retrieving Spoken Documents", School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213-3890.
5. Rabiner L.R, Fellow, IEEE, Stephen E Levinson, Member, IEEE " Isolated and Connected Word Recognition – Theory and Selected Applications ", IEEE Transactions on Communications, Vol Com 29 No. 5 May 1981.
6. Grott, R., & Schwartz, P. (2001, June),"Speech recognition from alpha to zulu", Paper presented at the Instructional Course in RESNA 2001 Conference, Reno, NV.
7. www.w3c.org/voice

8. Laver, J. (1994),"Principles of Phonetics", Cambridge: Cambridge University Press.

9. Lee, K, Giachin, E., Rabiner, R., L. P., and Rosenberg, "A. Improved Acoustic Modeling for Continuous Speech Recognition", DARPA Speech and Language Workshop. Morgan Kaufmann Publishers, San Mateo, CA, 1990.
10. Lee, K, "Automatic Speech Recognition: The Development of the Sphinx System". Kluwer Academic Publishers, Boston, 1989.
11. www.speech.cs.cmu.edu