Speech Recognition

Anjali Kalra[1], Sarbjeet Singh[2], Sukhvinder Singh[3]
M.Tech. CSE(1st Year).
Sri Sai College of Engg. And Technology, Pathankot[1,2,3]

Abstract:
Language is man's most important means of communication and speech its primary medium. Speech provides an international forum for communication among researchers in the disciplines that contribute to our understanding of the production, perception, processing, learning and use. Spoken interaction both between human interlocutors and between humans and machines is inescapably embedded in the laws and conditions of Communication, which comprise the encoding and decoding of meaning as well as the mere transmission of messages over an acoustical channel. Here we deal with this interaction between the man and machine through synthesis and recognition applications. The paper dwells on the **speech technology** and conversion of speech into analog and digital waveforms which is understood by the machines. Speech recognition, or speech-to-text, involves capturing and digitizing the sound waves, converting them to basic language units or phonemes, constructing words from phonemes, and contextually analyzing the words to ensure correct spelling for words that sound alike. **Speech Recognition** is the ability of a computer to recognize general, naturally flowing utterances from a wide variety of users. It recognizes the caller's answers to move along the flow of the call. We have emphasized on the modeling of speech units and grammar on the basis of **Hidden Markov Model**. Speech Recognition allows you to provide input to an application with your voice. The **applications and limitations** on this subject has enlightened us upon the impact of speech processing in our modern technical field. While there is still much room for improvement, current speech recognition systems have remarkable performance.

Key words:
Speech Technology, Hidden Markov Model

1. Introduction

One of the most important inventions of the nineteenth century was the telephone. Then at the midpoint of twentieth century, the invention of the digital computer amplified the power of our minds, enabled us to think and work more efficiently and made us more imaginative then we could ever have imagined .now several new technologies have empowered us to teach computers to talk to us in our native languages and to listen to us when we speak (**recognition**); haltingly computers have begun to understand what we say. Having given our computers both oral and aural abilities, we have been able to produce innumerable computer applications that further enhance our productivity. Such capabilities enable us to route phone calls automatically and to obtain and update computer based information by telephone, using a group of activities collectively referred to as Voice Processing.

2. Speech Technology

Three primary speech technologies are used in voice processing applications: stored speech, text-to – speech and speech recognition. Stored speech involves the production of computer speech from an actual human voice that is stored in a computer's memory and used in any of several ways.
Speech can also be synthesized from plain text in a process known as text-to – speech which also enables voice processing applications to read from textual database.
Speech recognition is the process of deriving either a textual transcription or some form of meaning from a spoken input. Speech analysis can be thought of as that part of voice processing that converts human speech to digital forms suitable for transmission or storage by computers. Speech synthesis functions are essentially the inverse of speech analysis – they reconvert speech data from a digital form to one that's similar to the original recording and suitable for playback. Speech analysis processes can also be referred to as a digital speech encoding (or simply coding) and speech synthesis can be referred to as Speech decoding.

3. Digitization of Analog Waveforms

Two processes are required to digitize an analog signal:
(a) Sampling, which discretizes the signal in time?
(b) Quantizing, which discretizes the signal in amplitude?

Manuscript received June 5, 2010
Manuscript revised June 20, 2010

3.1 Analysis/Synthesis in The Time and Frequency Domain

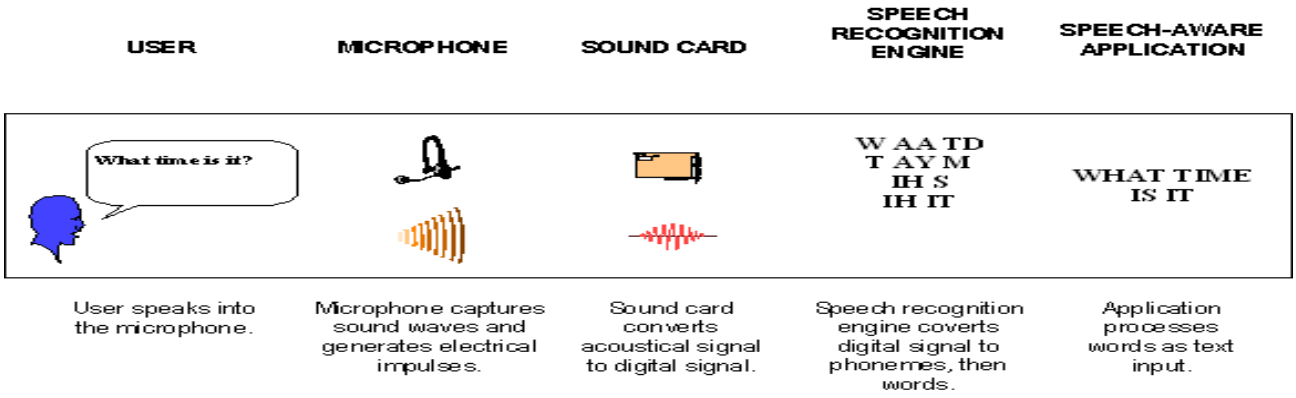
The analog and digital speech waveforms exist in time domain; the waveform represents speech as amplitude versus time. The time –domain sound pressure wave emanating from the lips is easily converted by microphone to a speech waveform, so it's natural that speech analysis/synthesis systems operate directly upon this waveform. The objective of every speech-coding scheme is to produce code of minimum data rate so that a synthesizer can reconstruct an accurate facsimile of the original speech waveform. Frequency domain coders attempt to reach this objective by exploiting the resonant characteristics of the vocal tract.

VOCODERS – Voice Coders
RELP – Residual–excited linear production
SBC – Subband Coding
CVSD- Continously variable slope deltmodulation
ADM - Adaptive Delta Modulation
ADPCM – Adaptive Quantization
LOG PCM – Logarithmic Quantization

Figure summarizes the relative voice quality of various speech coders. Voice quality is measured in terms of signal-to-noise ratio on the y-axis versus data rate as a logarithmic scale on the x-axis. Both solid and dashed traces appear in the figure and respectively represent objective and estimated results- all of which are approximations.

4. Speech Recognition

Speech recognition is the process of deriving either a textual transcription or some form of meaning from a spoken input. Speech recognition is the inverse process of synthesis, conversion of speech to text. The Speech recognition task is complex. This involves the computer taking the user's speech and interpreting what has been said. This allows the user to control the computer (or certain aspects of it) by voice, rather than having to use the mouse and keyboard, or alternatively just dictating the contents of a document. It would be complicated enough if every speaker pronounced every word in an identical manner each time, but this doesn't happen.



5. Variations in Speech

A Speech recognizer inputs a waveform, extracts information from it, uses that information to hypothesize words chosen from its vocabulary, applies knowledge of the grammar to prune or add word choices, and outputs tha recognized word or sequence of words. Speech recognizer must rely upon the waveform's record of speech events to accomplish their task, but the waveform also contains unwanted , irrelevant , or ambiguous information that works to confound the recognizer. The recognition process is further complicated because the production of phonemes and transitions between them is not uniform from person to person or from instance with the same talker. This lack of uniformity seriously complicates the task of automatic speech recognition.

5.1 Classifications of Systems

When a Speech recognition system requires words to be spoken individually, in isolation from other words, its said to be isolated-word system and recognizes only discrete words and only when they are separated from their neighbours by distinct interword pauses.
Continuous speech recognizing systems, allow a more fluent form of talking. Large-vocabulary sytems are defined to be those that have ore than one thousand words in their vocabularies; the others are considered small-vocabulary systems.
Finally, recognizers designed to perform with lower bandwidth waveforms as restricted by the telephone network are differentiated from those that require a broader bandwidth.

5.2 Speaker Dependence

An issue of utmost concern is whether or not a recognizer must be trained for the voice of each user. If so, the recognizer is said to be speaker-dependent, and it must know each talker's identity. Training consists of an enrollment procedure whereby the talker is prompted by the system to speak a number of words or sentences.

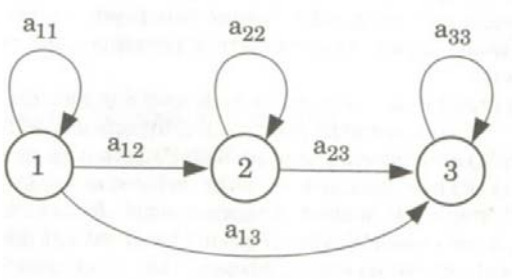
5.3 Vocabulary Size

Large vocabulary systems require their own recognition techniques. Because of the massive amounts of storage and processing power required, large vocabulary recognizers are not able to characterize and train each word individually as is done in small vocabulary systems. Large vocabulary systems adopt a different segment of speech to characterize, a segment smaller than words. Syllables, demisyllables, phonemes, and other units have been used.

5.4 Telephone Speech

Some recognizers, have been expressly designed for telephone operation. These are usually speaker-independent. Some allow unconstrained speech but might be restricted to only a few words, often the digits zero through ten and a few commands such as "yes" and "no." Some telephone recognizers are wordspotting systems that scan the incoming speech signal, trying only to pick out predefined keywords.

6. Hidden Markov Model

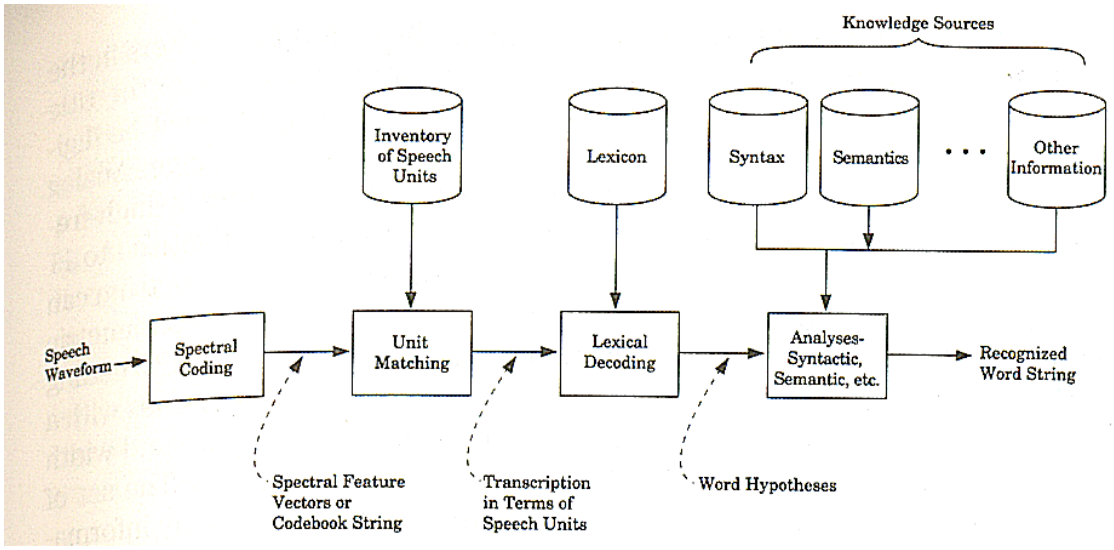


Hidden Markov Model : State Diagram

A hidden Markov model 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. It might be helpful to think of the unknown process as a black box about whose workings nothing is known except that, at each interval, it issues one member chosen from the alphabet. These models are called "hidden" Markov models precisely because the state sequence that produced the observable output is not known-it's "hidden." HMMs have been found to be especially apt for modeling speech processes.

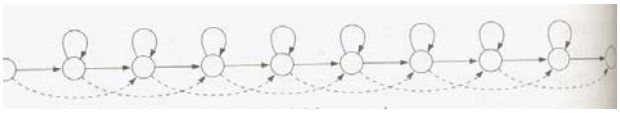
7. Choice of Speech Units

The amount of storage required and the amount of processing time for recognition are functions of the number of units in the inventory, so selection of the unit will have a significant impact. Another important consideration in selecting a speech unit concerns the ability to model contextual differences. Another consideration concerns the ease with which adequate training can be provided.



8. Modeling Speech Units with Hidden Markov Models

Suppose we want to design a word-based, isolated word recognizer using discrete hidden Markov models. Each word in the vocabulary is represented by an individual HMM, each with the same number of states. A Word can be modeled as a sequence of syllables, phonemes, or other speech sounds that have a temporal interpretation and can best be modeled with a left-to-right HMM whose states represent the speech sounds. Assume the longest word in the vocabulary can be represented by a 10-state HMM. So, using a 10-state HMM like that of Figure below for each word, let's assume states in the HMM represent phonemes. The dotted lines in the figure are null transitions, so any state can be omitted and some words modeled with fewer states. The duration of a phoneme is accommodated by having a state transition returning to the same state. Thus, at a clock time, a state may return to itself and may do so at as many clock times as required to correctly model the duration of that phoneme in the word, Except for beginning and end states, which represent transitions into and out of the word, each state in the word model has a self-transition. Assume, in our example, that the input speech waveform is coded into a string of spectral vectors, one occurring every 10 milliseconds, and that vector quantization further transforms each spectral vector to a single value that indexes a representative vector in the codebook. Each word in the vocabulary will be trained through a number of repetitions by one or more talkers. As each word is trained, the transitional and output probabilities of its HMM are adjusted to merge the latest word repetition into the rnodel. During training, the codebook is iterated with the objective of deriving one that’s optimum for the defined vocabulary. When an unknown spoken word is to be *recognized*, it's transformed to a string of code book indices. That string is then considered an HMM observation sequence by the recognizer that calculates , for each word model in the vocabulary, the probability of that HMM having generated the observations. The word corresponding to the word model with the highest probability is selected as the one recognized.



9. Acoustic/Phonetic Example using Hidden Markov Model

Every speech recognition system has its own architecture. Even those that are based on HMMs have their individual designs, but all share some basic concepts and features, many of which are recognizable even though the names are often different. A representative block diagram is given below. The input to a recognizer represented by Figure arrives from the left in the form of a speech waveform, and an output word or sequence of words emanates from the recognizer to the right.

It incorporates
(A) SPECTRAL CODING
The purpose of spectral coding is to transform the signal into a digital form embodying speech features that facilitate subsequent recognition tasks. In addition to spectral coding, this function is sometimes called spectrum analysis, acoustic parameterization, etc. Recognizers can work with time-domain coding, but spectrally coded parameters in the frequency domain have advantages and are widely used-hence the title "spectral coding."

(B) UNIT MATCHING
Its objective is to transcribe the output data stream from the spectral coding module into a sequence of speech units. The function of this module is also referred to as feature analysis, phonetic decoding, phonetic segmentation, phonetic processing, feature extraction, etc.

(C) LEXICAL DECODING
This module matches strings of speech units in the unit matching module's output stream with words from the recognizer's lexicon. It outputs candidate words-usually in the form of a word lattice containing sets of alternative word choices.

(D) SYNTACTIC, SEMANTIC, AND OTHER ANALYSES
Analyses that follow lexical decoding all have the purpose of pruning worst candidates passed along from the lexical decoding module until [mal word selections can be made. Various means and various sources of intelligence- can be applied to this end. Acoustic information (stress, intonation, change of amplitude or pitch, relative location of formants, etc.) obtained from the waveform can be employed, but sources of intelligence from outside the waveform are also available. These include syntactic, semantic, and pragmatic information.

10. Applications

10.1 Potential Applications for Speech Recognition

The specific use of speech recognition technology will depend on the application. Some target applications that

are good candidates for integrating speech recognition include:

10.1.1 Games and Edutainment

Speech recognition offers game and edutainment developers the potential to bring their applications to a new level of play. With games, for example, traditional computer-based characters could evolve into characters that the user can actually talk to.
While speech recognition enhances the realism and fun in many computer games, it also provides a useful alternative to keyboard-based control, and voice commands provide new freedom for the user in any sort of application, from entertainment to office productivity.

10.1.2 Data Entry

Applications that require users to keyboard paper-based data into the computer (such as database front-ends and spreadsheets) are good candidates for a speech recognition application. Reading data directly to the computer is much easier for most users and can significantly speed up data entry.
While speech recognition technology cannot effectively be used to enter names, it can enter numbers or items selected from a small (less than 100 items) list. Some recognizers can even handle spelling fairly well. If an application has fields with mutually exclusive data types (for example, one field allows "male" or "female", another is for age, and a third is for city), the speech recognition engine can process the command and automatically determine which field to fill in.

10.1.3 Document Editing

This is a scenario in which one or both modes of speech recognition could be used to dramatically improve productivity. Dictation would allow users to dictate entire documents without typing. Command and control would allow users to modify formatting or change views without using the mouse or keyboard. For example, a word processor might provide commands like "bold", "italic", "change to Times New Roman font", "use bullet list text style," and "use 18 point type." A paint package might have "select eraser" or "choose a wider brush."

11. Limitations

11.1 Speech Recognition

Each of the speech technologies of recognition and synthesis have their limitations. These limitations or constraints on speech recognition systems focus on the idea of variability. Overcoming the tendency for asr systems to assign completely different labels to speech signals which a human being would judge to be variants of the same signal has been a major stumbling block in developing the technology. The task has been viewed as one of de-sensitising recognisers to variability. It is not entirely clear that this idea models adequately the parallel process in human speech perception.
Human being are extremely good at spotting similarities between input signals - whether they are speech signals or some other kind of sensory input, like visual signals. The human being is essentially a pattern seeking device, attempting all the while to spot identity rather than difference.
By contrast traditional computer programming techniques make it relatively easy to spot differences, but surprisingly difficult to spot similarity even when the variability is only slight. Much effort is being devoted at the moment to developing techniques which can re-orientate this situation and turn the computer into an efficient pattern spotting device.

12. Merits

The uses of speech technology are wide ranging. Most effort at the moment centers around trying to provide voice input and output for information systems - say, over the telephone network.
A relatively new refinement here is the provision of speech systems for accessing distributed information of the kind presented on the Internet. The idea is to make this information available to people who do not have, or do not want to have, access to screens and keyboards. Essentially researchers are trying to harness the more natural use of speech as a means of direct access to systems which which more normally associated with the technological paraphernalia of computers.
Clearly a major use of the technology is to assist people who are disadvantaged in one way or another with respect to producing or perceiving normal speech.
The eavesdropping potential referred to in the slide is not sinister. It simply means the provision of, say, a speech recognition system for providing an input to a computer when the speaker has their hands engaged on some other task and cannot manipulate a keyboard - for example, a surgeon giving a running commentary on what he or she is doing. Another example might be a car mechanic on his or her back underneath a vehicle interrogating a stores computer as to the availability of a particular spare part.

13. Conclusion

Speech recognition is a truly amazing human capacity, especially when you consider that normal conversation requires the recognition of 10 to 15 phonemes per second.

It should be of little surprise then that attempts to make machine (computer) recognition systems have proven difficult. Despite these problems, a variety of systems are becoming available that achieve some success, usually by addressing one or two particular aspects of speech recognition. A variety of speech synthesis systems, on the other hand, have been available for some time now. Though limited in capabilities and generally lacking the "natural" quality of human speech, these systems are now a common component in our lives.

References

[1] W. Stokoe, D. Casterline, and C. Croneberg, A Dictionary of American Sign Language on Linguistic Principles, Gallaudet College Press, Washington D.C., USA, 1965.

[2] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," IEEE Trans. PAMI, vol. 27, no. 6, pp. 873–891, June 2005.

[3] T.S. Huang Y. Wu, "Vision-based gesture recognition: a review," in Gesture Workshop, Gif-sur-Yvette, France, 1999, vol. 1739 of LNCS, pp. 103–115.

[4] G. Yao, H. Yao, X. Liu, and F. Jiang, "Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm," in Intl. Conf. Pattern Recognition, Hong Kong, Aug. 2006, vol. 3, pp. 312–315.

[5] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of american sign language," Computer Vision & Image Understanding, vol. 81, no. 3, pp. 358–384, Mar. 2001.

[6] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in European Conf. Computer Vision, 2004, vol. 1, pp. 390–401.

[7] S. B. Wang, A. Quattoni, Louis-Philippe Morency, David Demirdjian, and Trevor Darrell, "Hidden conditional random fields for gesture recognition," in Computer Vision & Pattern Recognition, New York, USA, June 2006, vol. 2, pp. 1521–1527.

[8] J. L¨o¨of, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in ICSLP, Pittsburgh, PA, USA, Sept. 2006.

[9] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, "Deformation models for image recognition," IEEE Trans. PAMI, p. to appear, 2007.

[10] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, "Tracking using dynamic programming for appearance-based sign language recognition," in IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Southampton, Apr. 2006, pp. 293–298.

[11] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee, The Syntax of American Sign Language, MIT Press, 1999.

[12] T. K¨olsch, D. Keysers, H. Ney, and R. Paredes, "Enhancements for local feature based image classification," in Intl. Conf. Pattern Recognition, Cambridge, UK, Aug. 2004, vol. 1, pp. 248–251.

[13] A. Zolnay, R. Schl¨uter, and H. Ney, "Acoustic feature combination for robust speech recognition," in ICASSP, Philadelphia, PA, Mar. 2005, vol. 1, pp. 457–460.

[14] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," Speech Communication, vol. 38, pp. 19–28, 2002.

[15] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," IEEE Trans. PAMI, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[16] P. Dreuw, D. Stein, and H. Ney, "Enhancing a sign language translation system with vision-based features," in Intl. Workshop on Gesture in HCI and Simulation 2007, Lisbon, Portugal, May 2007, p. to appear.

[17] Dillon, T.W., & Norcio, A. F. (1997, October). User performance and acceptance of a speech-input interface in a health assessment task. International Journal of Human-Computer Studies, 47(4), 591-602.