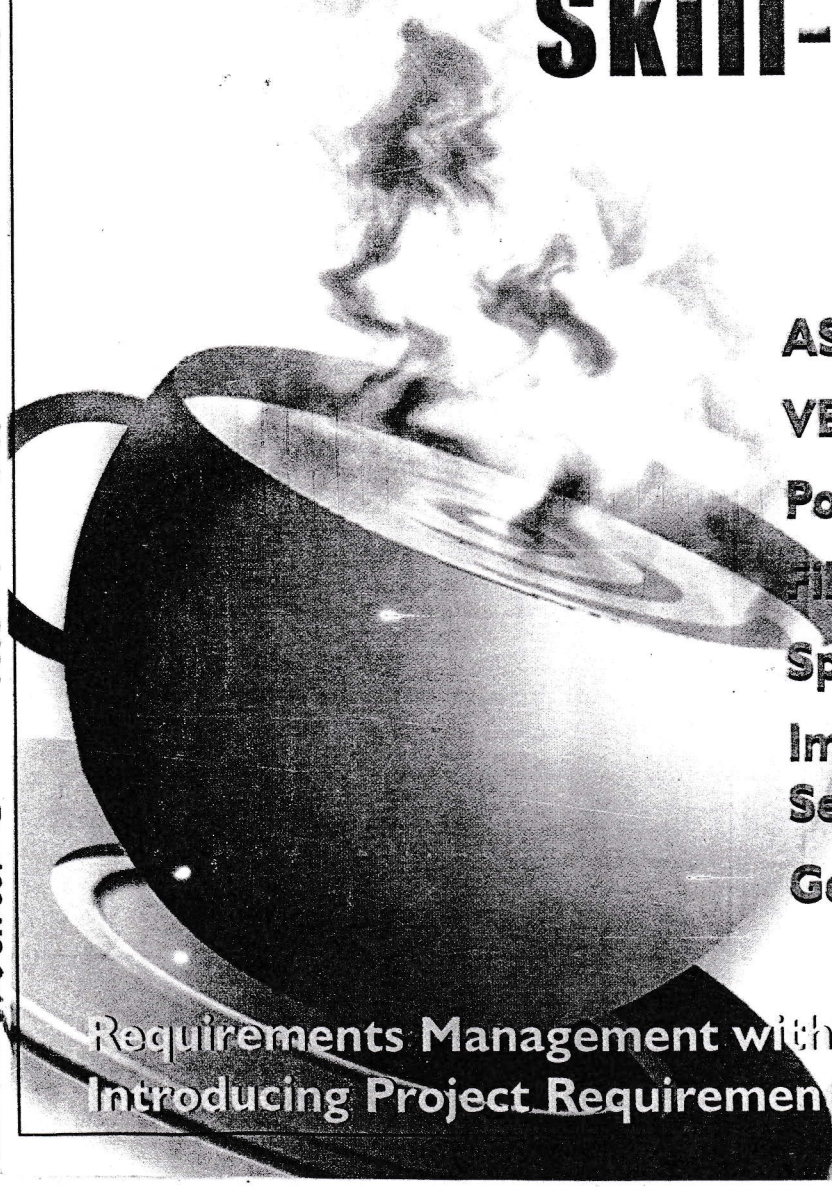
Asia's # 1 Software Technology Magazine

DeveloperIQ

2 Free CDs

US $ 15 Rs. 100 Vol. 4 No. 12, December 2004

HOTTEST Skill-Sets 2005



ASP .Net 2.0

VB.Net Tips

Porting using Winelib

File Objects in Python

Speech Recognition

Implementing MVC using Servlets, Javabeans & JSP

Getting started with RSS feeds

Requirements Management within the CMMI

Introducing Project Requirements

An TECHMEDIA Publication

Challenges in Automatic Speech Recognition

— Mohammad Saleem

Speech recognition is an important area of research because it offers an improved form of human-machine interaction. The author discusses the problems that need to be overcome to build accurate speech recognition systems.

Speech recognition continues to be a challenging field for researchers. The successful speech recognizer is free from the constraints of speakers, vocabularies, ambiguities and environment. A lot of efforts have been made in this direction, but complete accuracy is still far from reach. The task is not an easy one due to the interdisciplinary nature of the problem.

The recognition score degrades due to significant variations in the way different words of a language are pronounced by users from different regions and dialectal backgrounds. Also the variability of the voice characteristics from one speaker to another affects the recognition accuracy. The transducer, environment and reverberation are other factors. This article reviews challenges to automatic speech recognition, which are major concerns for scientists.

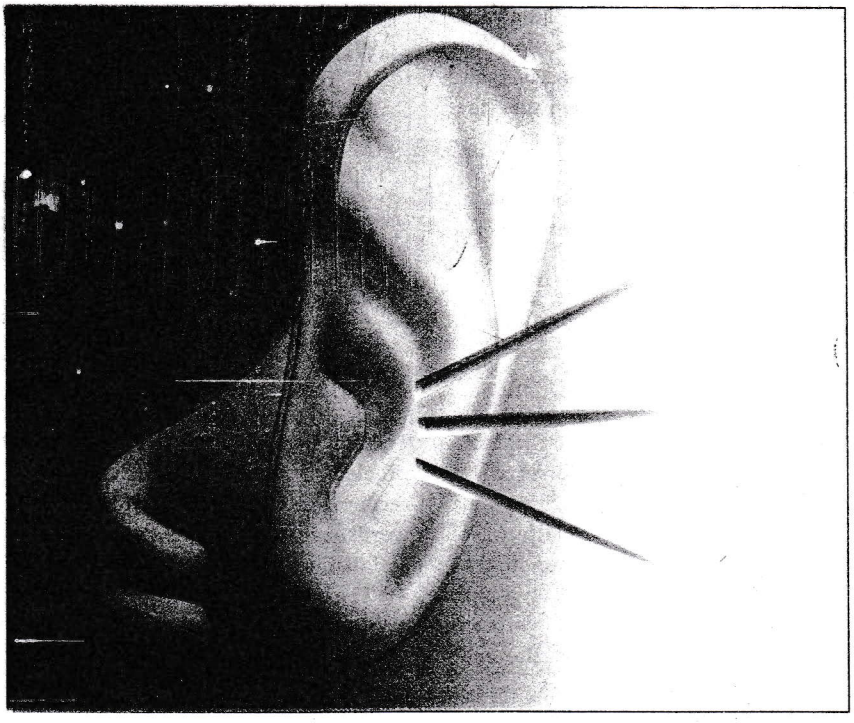
Introduction

Speech Recognition got its first jump-start in AT&T's Bell Labs in 1936 when researchers developed the first electronic speech synthesizer. After more than three decades, Threshold Technology introduced the first commercial speech recognition product in the early 1970's - the VIP100 system. While the system performed some speech recognition, it could only recognize a very limited number of discrete-words from a specific user.

The problem with these early efforts was that they focused primarily on the end product of speech - the words people were attempting to generate. By the mid-1980s, speech recognition programmers were building on Chomsky's ideas of grammatical structure, and using more powerful hardware to implement statistical phoneme-chain recognition routines. It was not until the dramatic increase of processing power of the 1980's and 1990's that technology companies could seriously consider the possibility of speech recognition. Further, many of the major companies collaborated to create application-programming standards. This millennium has brought a wide array of new speech recognition products and research areas.

Automatic Speech Recognition

For decades, people have been dreaming of an "intelligent machine" which can master natural speech. Automatic Speech Recognition (ASR) is one subsystem of this 'machine', the other subsystem being Speech Understanding (SU). The goal of ASR is to transcribe natural speech while



the goal of SU is to understand the meaning of transcription.

Automatic Speech Recognition means different things to different people. At one end of the spectrum is the voice-operated alarm clock which ceases ringing when the word 'stop' is shouted at it, and at the other end is the automatic dictating machine which produces a typed manuscript in response to the human voice or the expert system which provides answers to spoken questions. A practical speech recognizer falls somewhere between these two extremes.

Disciplines Involved in ASR

Automatic recognition of speech has been a goal of research for more than four decades. Its interdisciplinary nature adds to the difficulty of performing research in this area. Tendency of most researchers is to apply a monolithic approach to individual problems.

- **Signal Processing** - The process of extracting relevant information from the speech signal in an efficient and robust manner.
- **Physics (acoustics)** - The science of understanding the relationship between the physical speech signal and physiological mechanisms (the human vocal tract mechanism).
- **Pattern Recognition** - The set of algorithms used to cluster data to create one or more prototypical patterns and to match a pair of patterns.
- **Communication** & Information Theory - The procedures for estimating parameters of statistical models; the methods for detecting the presence of particular speech patterns, the set of modern coding and decoding algorithms.
- **Linguistics** - The relationship between sounds (phonology), words in a language (syntax), meaning of spoken words (semantics), and sense derived from meaning (pragmatics).
- **Physiology** - Understanding the higher order mechanisms within the human central nervous system that account for speech production and perception in human beings.
- **Computer Science** - The study of efficient algorithms for implementing ASR in software or hardware.
- **Psychology** - The science of understanding the factors that enables a technology to be used by human beings in practical tasks.

Problems of ASR

Significant progress in ASR has been achieved by increasing the awareness of the problems of speech recognition and the application of various techniques to attempt to solve these problems.

At one end of the spectrum is the voice-operated alarm clock which ceases ringing when the word 'stop' is shouted at it, and at the other end is the automatic dictating machine

A. Recognition Units

The first step in solving the recognition problem is to decide which units to recognize. Possible candidates are words, syllables, diphones, phonemes and distinctive features.

- **The Words** - The word is a basic unit in speech recognition. The meaning of an utterance can only be deduced after the words of which it is composed have been recognized. The word is so basic that a whole class of speech recognizers, discrete utterance or isolated word recognizers has been designed for the sole purpose of identifying spoken words. However, these devices require each word to be spoken in isolation; this is not the way in which sentences are normally produced.

One of the problems of employing the word as the recognition unit is the number of words in the language. This is of the order of 100,000. For recognition to take place some representation of each word needs to be stored. This implies that a large, though not impossible, amount of storage is required. Nevertheless there are a number of domains of man-machine interaction where the vocabulary can be restricted to a much smaller number of words. In these situations the word is often employed as the recognition unit.

Another problem encountered in using words as the recognition units is in determining where one word ends and the next begins. There is often no acoustic evidence of word boundaries. In fact, in the pronunciation of continuous speech, co-articulation effects take place across word boundaries, altering the acoustic manifestation of each word, and obscuring the boundaries.

- **The Syllables** - Instead of using words as the recognition units, smaller units such as the syllable may be considered. The syllable is attractive as it has a fixed structure. It consists of an initial consonant or consonant cluster, a medial vowel or diphthong, and a final consonant or consonant cluster, C_iVC_f. The vowel is obligatory but the consonants are optional. The intensity of every consonant is less than that of the vowels, so an algorithm can be devised for segmenting the speech stream into

syllables. Problems arise, however, with strings of consonants, as it is often difficult to decide whether a consonant is part of the final consonant cluster of the last syllable or part of the initial consonant cluster of the next syllable.

- **The Demisyllables** - A very significant reduction in the number of units can be achieved by employing demisyllables instead of syllables. A demisyllable consists of half a syllable, from the beginning of the syllable to the middle of the vowel, C_iV, or from the middle of the vowel to the end of the syllable, VC_f. A syllable can be segmented into demisyllables by splitting it at the point of maximum intensity.
- **The Diphones** - Another possible recognition unit is the diphone. These have been found useful as the unit for speech synthesis but the number required, 1000-2000, is similar to that of demisyllables. The problem of segmentation, deciding where one ends and the next begins, however, is much more difficult with diphones than demisyllables.
- **The Phoneme** - The other possible recognition unit is the phoneme. The advantage of the phoneme is the small number (Approx. 40-60 phoneme). However, phonemes have a number of contextual variations known as allophones, and there are some 100-200 of these. Even so the small numbers involved make the phoneme, or phone, an attractive recognition unit.

The problem with phoneme recognition units is segmentation. Co-articulation effects modify the acoustic manifestation of each phoneme. Except in certain cases where a voiced phoneme is followed by a voiceless one, or vice versa, it is impossible to tell where one phoneme ends and the next begins.

8. Variability

There are a great number of factors, which cause variability in speech. These include the speaker, the context, the speaking rate, the environment (extraneous sound and vibration) and the transducer employed.

- **The Speaker** - The speech signal is very dependent on the physical characteristics of the speaker. The size of the vocal tract increases during childhood, and this gives rise to different formant frequencies for the productions of the same vowel at different ages. The vocal tracts of men are, on average, about 30% longer than those of women. This again gives rise to different formant frequencies. Age and sex of the speaker cause great variations in fundamental frequencies of speech sounds.

People from different parts of the country and different social and economic backgrounds speak with different dialects

People from different parts of the country and different social and economic backgrounds speak with different dialects. This variability is even greater with people speaking a second language. The competence with which it is spoken depends on the motivation, intelligence, and perceptual and motor skills of the speaker, and also on the age at which the second language was learned.

Even the same speaker uttering the same words on different occasions shows some variability. When a person first encounters a speech recognizer he will be in an unfamiliar situation and so will speak to it in a formal manner. However, it is on this occasion that the machine will be trained to recognize his voice. At subsequent meetings he will be more relaxed, so he will address the machine in a less formal manner. This may cause the recognition accuracy to decline.

- **The Context** - The production of each word still exhibits variability, even when a familiar situation has been reached. Co-articulation effects cause each word to be pronounced differently depending upon context. The articulators anticipate the beginning of the next word whilst the end of the present word is still being produced. Words are also pronounced differently depending on their position in the sentence and their degree of stress.
- **The Speaking Rate** - Another source of variability is speaking rate. The tempo of speech varies widely depending upon the situation, the topic being discussed and the emotional state of the speaker. Unfortunately the duration of all sounds in fast speech is not reduced proportionally compared with their duration in slow speech. In fast speech, pauses are eliminated and steady sounds, such as vowels, are compressed whilst the duration of some consonants remain almost constant.

The amplitude of speech signal depends on the amount of vocal effort employed and the distance of the microphone from the mouth. The voc

effort affects the shape of the glottal pulse, and thus the intensity and frequency of the speech signal. The distance between the microphone and the mouth can vary with a hand-held microphone, but can be kept approximately constant by means of a microphone on a boom attached to a headset.

- **The Environment** - The background sound in many circumstances is an uncontrollable variable. If this sound is constant and always present, such as the hum from the cooling fan in a computer, the level can be measured and its effect subtracted from the speech signals. If the background noise level is variable, it is important that the signal should be made as high as possible. This is usually achieved by holding the microphone close to the mouth, and by using a directional, noise-canceling microphone.
- **The Reverberation** - The speech signal may be distorted by reverberation. As well as the direct path from the mouth to the microphone, there will be other acoustic paths due to reflections from objects such as walls and furniture. These paths will be longer than the direct path, and so will add delayed and distorted versions of the signal to the original. Working in an anechoic chamber could eliminate reverberation, but this is not usually practical. It should be noted that the introduction of extra items of equipment or the presence of other bodies might distort the signal.
- **The Transducer** - The transducer, used for converting the acoustic signal into an electrical signal may introduce distortion. If the same microphone is always used, this will not cause variability. If different microphones are used, however, as will be the case when a speech recognizer is used via the telephone system, the characteristics of the different microphones and their associated transmission channels, will introduce variability.

C. Ambiguity

A further problem for a speech recognizer is that of ambiguity. This becomes important when the system is required to perform some action as a result of the signals, which it has received.

- **Homophones** - There are a number of words, which have different spellings, and meanings, but which, nevertheless, sound alike. For example, consider the words 'to', 'too' and 'two'. In applications such as a speech-driven word processor, homophones present problems. These problems cannot be resolved at the acoustic or phonetic levels. Recourse must be had to higher levels of linguistic analysis.
- **Overlapping Classes** - The first and second formant frequencies (F1 & F2) are able to identify most of vowels. Though, if we plot a graph of F1 versus F2 for a large number of vowels spoken by a variety of speakers, the points plotted do not form separate areas. For e.g., the vowel /a/ spoken by one person may have identical formant frequencies to a vowel /ɔ/ spoken by another.
- **Syntactic Ambiguity** - Even if the phoneme sequence can be recognized and correctly segmented into words, there may still be ambiguity of meaning until all the words are grouped into appropriate syntactic units.
- **Word boundaries** - Another problem of ambiguity concerns the location of word boundaries. Occasionally a sequence of phonemes occurs which has one interpretation with the word boundary inserted at one location, and another meaning with it inserted at another location. This may involve shifting the boundary by a single phoneme, such as /greɪtɛɪp/ which may be interpreted as 'grey tape' or 'great ape', or it may mean moving the word boundary by a whole syllable, for example /laɪthauskiːpə/ may mean 'light housekeeper' or 'lighthouse keeper'.

Relevant issues of ASR design	
Environment	Type of Noise, Signal/Noise Ratio; Working Conditions
Transducer	Microphone, Telephone
Channel	Band Amplitude; distortion, echo
Speakers	Speaker-dependence, speaker-independence; sex; age; physical and psychical state
Speech Styles	Voice Tone (quiet, normal, shout) production (isolated words, continuous speech, read or spontaneous speech); speed (slow, normal, fast)
Vocabulary	Characteristics of available training data; specific or generic vocabulary

Classification of ASR Systems

Vocabulary Size Restrictions

- Small: 100-300 words
- Medium: 1000 words

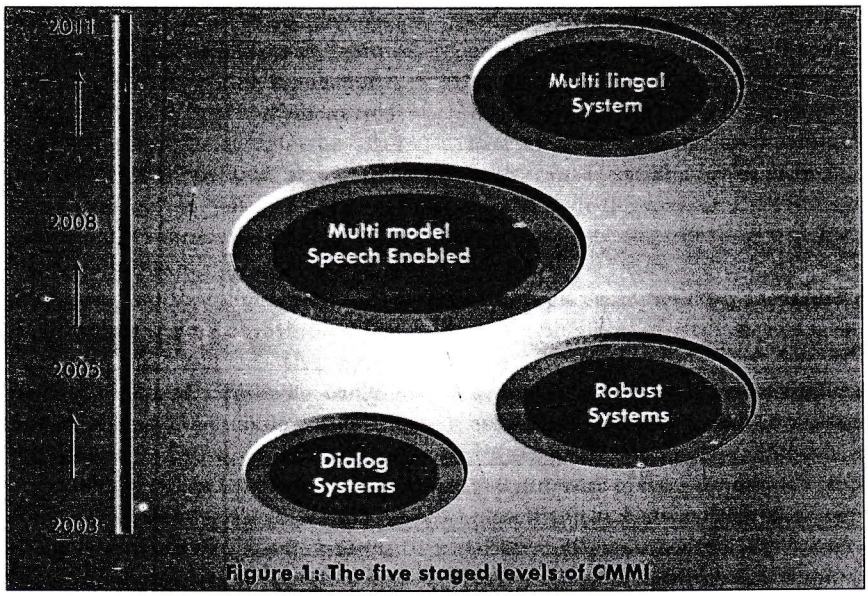


Figure 1: The five staged levels of CMMI

- Large: 10k-50k words

Speaking Style Restrictions

- Isolated Word
- Connected Word
- Continuous Word

Speaker Dependence

- Speaker-dependent
- Speaker-adaptive
- Speaker-independent (Closed-speaker)
- Speaker-independent (Open-speaker)

Environment Dependence

- Studio
- Ordinary Room
- Noisy Room

The Future Challenges of ASR

Looking ahead, the ultimate challenge for designers is to develop a system that can match the ability of humans to recognize languages. According to Microsoft's Whisper project on speech research, their goal is to "develop a general purpose, speaker-independent continuous speech recognition engine that can recognize unrestricted text and is effective for command and control, dictation, and conversational systems." While this ambitious goal appears lofty, it may not be that far away.

Future research will continue to improve statistical models for analyzing speech, not only by improving mathematical algorithms that adapt to the unique style of the speaker, but also by better control of the varying environments from which speaker might use the application. Researchers are expending a fair amount of resources to strengthen the underlying statistics behind their language models. However, speech recognition devices will never proliferate the mainstream market unless they are able to offer better control of outside noise (e.g. in an office or in a car).

Conclusion

We have discussed the problems in building accurate and robust speech recognition systems. We categorize these problems as recognition units (phoneme, syllables, diphones etc.), variability (speaker, context, environment etc.) and ambiguity (homophones, word boundaries etc.). The basic problem is the paradox that speech consists of a continuous stream of sound with no obvious discontinuities at the boundaries between the words, and yet speech is perceived as a sequence of words. It is almost impossible to predict accurately the rate of progress in any scientific field. However, based on the rate of progress over the past decade, it seems reasonable to make some broad projections as to where speech recognition is headed in the next decade (**Figure-1**). DIQ

References

1. Ainsworth, A. William, "Speech Recognition by Machine", IEEE Computing Series 12, 1988
2. Becchetti C, Ricotti L. R., "Speech Recognition - Theory & C++ implementation", John Wiley & Sons Ltd, 1999
3. Chris Rowden, "Speech Processing", McGraw Hill Publication, 1992
4. Joseph Picone, "Signal Modelling Techniques in Speech Recognition", Proceedings IEEE, 1993
5. Junqua JC, Haton JP, "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1996
6. Rabiner L, Juang B., "Fundamentals of Speech Recognition", Prentice Hall Signal Processing Series, 1993
7. Wai, C. Chu, "Speech Coding Algorithms", John Willey & Sons, 2003
8. Speech Software Technology (I) Pvt. Ltd, www.sstil.com/sst_technology.htm

The author is a Ph.D student at the Dept. of Computer Science, Jamia Millia Islamia, New Delhi.
He can be reached on saleem_amity@rediffmail.com.