SPEECH RECOGNITION BY MACHINES

Bishnu S. Atal
AT&T Bell Laboratories, Murray Hill, NJ, U.S.A.

ABSTRACT

Voice communication with machines is no longer a dream but a reality. The tremendous progress that has been accomplished has come about as a result of solving successfully some of the fundamental problems caused by the immense variability present in the speech signal. This paper will discuss important issues in automatic recognition of speech, the major advances made in this field, the current state of the technology, and future developments.

INTRODUCTION

Speech is a natural form of communication for humans and thus the problem of understanding or "recognizing" speech by machines has challenged scientists for many years. We do not yet understand in any detail how humans understand speech. But, considerable advances in automatic speech recognition and understanding by machines have taken place [1-2].

Research in automatic speech recognition (ASR) since the 1970s has produced solutions for increasingly difficult tasks, from the correct recognition of a few isolated words from a single speaker to recognition of fluent speech from virtually any speaker. Progress in automatic recognition of speech is continuing and the research frontiers are shifting towards the solution of an even harder problem -- unconstrained dialogue with machines. The availability of high-speed processors and high density memories at reasonable cost in digital computers, along with large databases of recorded speech, has made it possible to develop sophisticated signal representations, pattern-matching techniques, and language models in support of automatic speech recognition.

COPING WITH ACOUSTIC VARIABILITY

Speech is the acoustic form of language. Speech recognition is essentially a process of recognizing acoustic patterns of the spoken language. Human communication by voice appears to be so simple that we often forget how variable these acoustic patterns are. Vast differences occur in the spoken utterance dependent on context, speaking style, speaker, dialect, speaking environment, microphone characteristics, etc. The major obstacle to achieving high accuracy in speech recognition is the large variability present in the speech signal, with only a small part that is important for carrying the linguistic information. A large part of this variability is due to various redundancies introduced by the human speech production process to achieve reliable speech communication in noisy and reverberant acoustic environments. Other sources of variability are introduced by differences in the vocal systems of speakers, differences in speaking rates, and the influence of neighboring sounds on the acoustic realization of a particular sound due to sluggish articulatory movements. Automatic methods of speech recognition must be able to handle this large variability in a proper manner. We illustrate here a few examples of the variability inherent in the speech signal.

Consider a simple case of the same word spoken by the same speaker on two different occasions. Acoustics realizations of the two utterances are in general not identical due to variations in the speaking rate or the speaking style. The

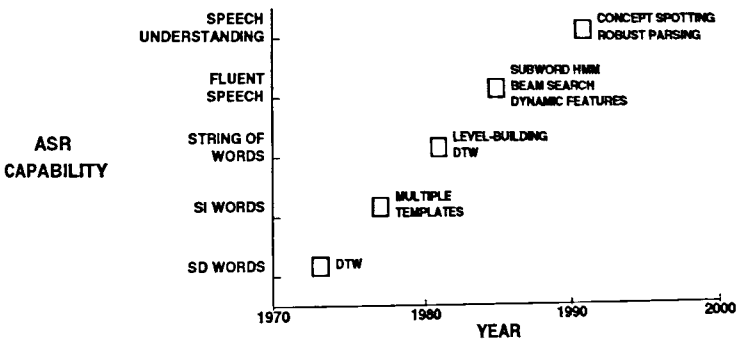

Figure 1. Major advances in automatic speech recognition from 1970 onwards that made it possible to achieve steady progress, from the simple problem of speaker-dependent recognition of isolated words to the more difficult problem of recognizing spontaneous speech encountered in dialogues.

speech recognition procedure must be able to compare the two utterances and conclude that they are same. Variations in speaking rate cause nonlinear distortions of the time axis of speech patterns. Linear scaling of the time axis is generally not sufficient to cope with speaking-rate variability. Nonlinear time normalization using dynamic programming is necessary to achieve time alignment between unknown and reference utterances.

The same word spoken by two different speakers will in general have different acoustic characteristics, due to the differences in their vocal tracts and speaking styles. A speaker-dependent recognizer uses the utterance of a single speaker to learn the speech patterns of that speaker. In contrast, a speaker-independent recognizer is trained on speech from many speakers and is used to recognize speech from speakers that may be outside of the training population.

Recognition of continuous speech introduces additional problems. In isolated words or speech where words are separated by distinct pauses, the beginnings and ends of words are clearly marked. In continuous speech, word boundaries are blurred and words evolve smoothly in time with no acoustic separation. Automatic methods of segmenting continuous speech into words therefore had to be devised. Machine recognition of continuous speech with a large vocabulary requires that syntactic and semantic constraints be incorporated in the recognition process.

The progress in automatic speech recognition has come about as a result of solving successfully problems created by the large variability present in speech signals. Figure 1 shows some of the major advances that were made during the past twenty years and were important in achieving this progress. Dynamic time warping (DTW) was the most important step taken in early 1970s to handle variations in speaking rates [3]. Clustering of speech patterns into multiple templates [4] for each word made it possible to recognize words spoken by any speaker (speaker-independent speech recognition). Recognition of individual words from a string of connected words required development of level-building

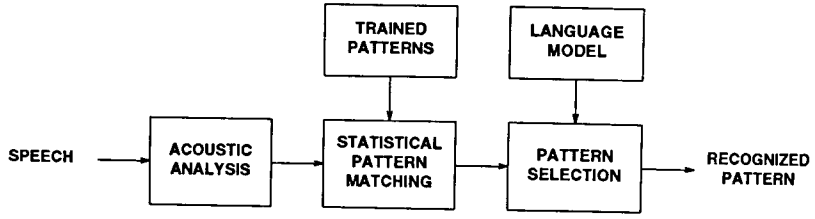

Figure 2. A block diagram showing the basic steps of the pattern recognition approach.

ASR algorithms [5]. The introduction of Hidden Markov Models (HMM) techniques [6] in early 1980s was a giant step that set the stage for big success in ASR. Recognition of continuous fluent speech was made possible by breaking words into phone-like subword units, use of bigram and higher-order language models, and the development of efficient beam search techniques [7]. Finally, introduction of concept spotting [8] and robust parsing techniques moved ASR from simple recognition of words to understanding meaning conveyed by a group of words.

AUTOMATIC SPEECH RECOGNITION PROCESS

There are at present three principal approaches to speech recognition: The first is based on statistical techniques of pattern recognition [6] that utilizes a training set of speech data to learn important information about the speech signal. The second approach, commonly known as acoustic-phonetic approach, uses knowledge of the relationship between acoustic and phonetic structures of the language. The third approach uses artificial neural networks. Best results in speech recognition have so far been achieved by using the statistical pattern recognition approach supplemented by the knowledge of acoustic-phonetic relationships in speech.

The basic steps of the pattern recognition approach are illustrated in the block diagram of Fig. 2. The speech signal is analyzed to provide a parametric representation at the acoustic level. These parameters ("features") are then compared to a stored set of patterns derived from a large collection of speech utterances from many speakers using a HMM-based training procedure. This comparison provides a set of scores representing the similarity between the unknown pattern and each of the stored patterns. The last step augments these scores with other knowledge about the speech utterance, such as the language, the context, and semantics, to yield the best recognition results.

Acoustic Features of Speech

The selection of proper acoustic or spectral features is crucial for achieving high performance in speech recognition. The short-time spectral envelope of speech, obtained either by filtering or linear prediction analysis, is still considered to be the most effective representation for speech recognition, especially if rendered on a critical-band ("Bark") scale. The spectra are computed sequentially in time at intervals of 10 to 20 ms and are usually converted into cepstral coefficients. The cepstrum is defined as the inverse Fourier transform of the logarithm of the power spectrum.

The cepstral coefficients are instantaneous (static) features. One of the most important advances in the acoustic representation of speech has been the introduction of dynamic features [9], such as first- and second-order derivatives of the

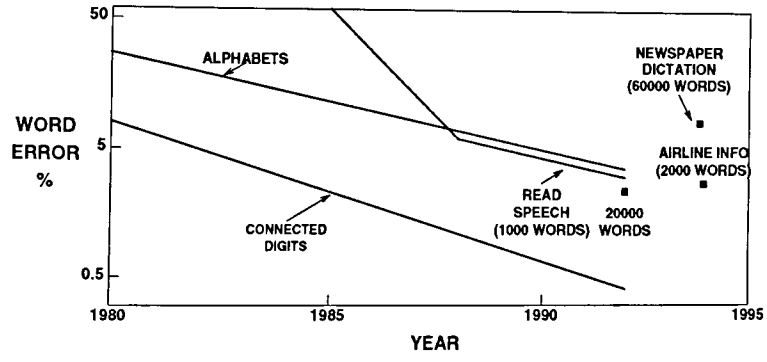

Figure 3. Reduction in the word error rate for different automatic speech recognition tasks between 1980 and 1995.

cepstrum. The static and dynamic features are generally combined to form a larger feature set; a smaller set can then be obtained by proper "pruning" from the larger set.

Training and Pattern Matching

Because of its statistical nature and its simple algorithmic structure for handling the large variability in speech signals, the HMM approach has found widespread use for automatic speech recognition [10]. Most of the successful systems today are based on this approach. An HMM representation can be used to model a sound, a word, a phrase, or a long utterance. During the training phase, the HMMs are trained from an ensemble of observation vectors coming from spoken utterances and stored for each of the basic pattern to be recognized. During recognition, the unknown acoustic patterns are compared with a set of stored reference patterns ("templates") established from the training data to provide a set of similarity scores between the test and reference patterns.

Pattern Selection

In the recognition phase, the utterance is decoded by determining the optimal sequence of HMM states and the corresponding speech units based on the observed sequence of acoustic feature vectors in the utterance. Search procedures based on dynamic programming methods are used to find the sequence of states with the maximum likelihood. Additional information based on the syntax and semantics of the source language is included in the recognition process to produce admissible outputs.

CURRENT CAPABILITIES

The performance of ASR systems continues to improve steadily. Figure 3 shows the word error rate for various test materials and the steady decrease in the error rate achieved since 1980. The performance of current ASR systems degrades considerably in the presence of noise, reverberation, or distortion and for conversational speech.

There are many factors that influence the performance of automatic speech recognition systems. The most important of these are the size of the vocabulary and the speaking style. Figure 4 shows examples of ASR tasks that can be handled by automatic methods for different vocabulary sizes and speaking styles. Generally, the number of confused words increases with the vocabulary size. Current systems can properly recognize a

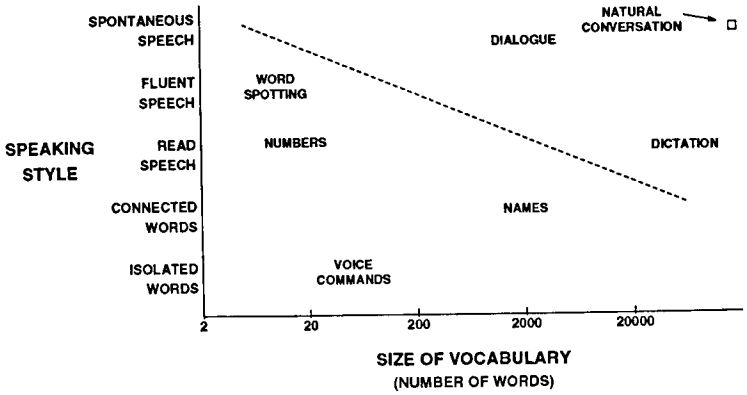

Figure 4. Different speech recognition tasks shown in a space of two dimensions: speaking style and size of vocabulary. Tasks that can be handled by current technology are shown to the left of the diagonal line. Items to the right of the line need more research to bring performance to a useful level.

vocabulary of as many as a several thousand words, while the speaking style can vary over a wide range, from distinct isolated words to spontaneous running speech.

Examples of speech recognition applications that can be handled by the current technology are shown on the left side of the diagonal line in Fig. 4. These include recognition of voice commands for computers, names, digit strings, and keyword spotting. The items on the right of the diagonal line in Fig. 4 are examples of the speech recognition tasks that work well in laboratory environments but which need more research to become useful for real-world applications. Automatic recognition of fluent speech with a large vocabulary is not feasible unless constraints on syntax and semantics are introduced. The present capability in handling natural languages and in following a dialogue is limited because we do not understand how to model the variety of expressions that natural languages use to convey concepts and meanings.

CHALLENGING ISSUES IN SPEECH RESEARCH

For speech technology to be used widely, it is necessary that the major roadblocks faced by the current technology be removed. Some of the key issues that pose major challenges in speech research are listed below:

- *Robust performance*. Can the recognizer work well for different speakers and in the presence of the noise, reverberation, and spectral distortion that are often present in real communication channels?
- *Automatic learning of new words and sounds*. In real applications, the users will often speak words or sounds that are not in the vocabulary of the recognizer. Can it learn to recognize such new words or sounds automatically?
- *Grammar for spoken language*. The grammar for spoken language is quite different from that which is used in carefully constructed written text. How does the system learn this grammar?
- *Flexibility*. Unless it is flexible, speech technology will have limited applications. What restrictions are there on the vocabulary? Can it handle spontaneous speech and natural spoken language?

A number of methods have been proposed to deal with the problem of robustness. The proposed methods include signal enhancement, noise compensation, spectral equalization, robust distortion measures, and novel speech representations. These methods provide partial answers valid for specific situations, but do not provide a satisfactory answer to the problem. Clean, carefully articulated, fluent speech is highly redundant, with the signal carrying significantly more information than is necessary to recognize words with high accuracy. However, the challenge is to realize the highest possible accuracy when the signal is corrupted with noise or other distortions and part of the information is lost.

FUTURE DEVELOPMENTS

The advances in digital technology is rapidly changing the fabric of telecommunications and the way we access information. A new mode of interacting with computers through voice is emerging. When combined with video, the voice mode offers an easy natural communication interface with computers. Speech recognition technology is a key component of such an interface. Human speech communication is a complex process and it will require scientific understanding of many other issues beyond acoustics and pattern recognition to mimic this process in computers. Speech science is expanding its frontiers to answer the basic question of how we put words together to express ideas in the spoken language.

REFERENCES

[1] Rabiner, L. and Juang, B.-H. (1993), *Fundamentals of speech recognition*, Englewood Cliffs: Prentice Hall.

[2] Makhoul, J. and Schwartz, R. (1994), "State of the art in continuous speech recognition", in D. Roe and J. Wilpon (eds.), *Voice communication between humans and machines*, Washington: National Academy Press, pp. 165-198.

[3] Itakura, F. (1975), "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. ASSP*, vol. ASSP-23, pp. 57-72.

[4] Rabiner, L. Levinson, S., Rosenberg, A., Wilpon, J. (1979), "Speaker independent recognition of isolated words using clustering techniques", *IEEE Trans. ASSP*, vol. ASSP-27, pp. 336-349.

[5] Meyers, C. and Rabiner, L. (1981), "A level-building dynamic time warping algorithm for connected word recognition", *IEEE Trans. ASSP*, vol. ASSP-29, pp. 284-297.

[6] Bahl, L., Jelenik, F., and Mercer, R. (1983), "A maximum likelihood approach to continuous speech recognition", *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-5, pp. 179-190.

[7] Chow, Y. et al. (1987), "BYBLOS: The BBN continuous speech recognition system", *Proc. IEEE-ICASSP*, Dallas, TX, pp. 89-92.

[8] Pieraccini, R. and Levin, E. (1992), "Stochastic representation of semantic structure for speech understanding", *Speech Communications*, vol. 11, pp. 283-288.

[9] Furui, S. [1986], "Speaker-independent isolated word recognition using dynamic features of speech

ICPhS 95 Stockholm Session 93.4 Vol. 1 Page 493

spectrum", *IEEE Trans. ASSP*, vol. ASSP-34, pp. 52-59.

[10] Rabiner, L. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-286.