

Speech Recognition, Automatic: History

B-H Juang, Georgia Institute of Technology, Atlanta, GA, USA

L R Rabiner, Rutgers University, New Brunswick, NJ, USA, and University of California, Santa Barbara, CA, USA

© 2006 Elsevier Ltd. All rights reserved.

Introduction

Speech is the primary means of communication between people. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities, to the desire to automate simple tasks inherently requiring human-machine interactions, research in automatic speech recognition (and speech synthesis) by machine has attracted a great deal of attention over the past five decades.

The desire for automation of simple tasks is not a modern phenomenon, but one that goes back more than 100 years in history. By way of example, in 1881 Alexander Graham Bell, his cousin Chichester Bell, and Charles Sumner Tainter invented a recording device that used a rotating cylinder with a wax coating on which up-and-down grooves could be cut by a stylus, which responded to incoming sound pressure (in much the same way as a microphone that Bell invented earlier for use with the telephone). Based on this invention, Bell and Tainter formed the Volta Graphophone Co. in 1888 in order to manufacture machines for the recording and reproduction of sound in office environments. The American Graphophone Co., which later became the Columbia Graphophone Co., acquired the patent in 1907 and trademarked the term 'Dictaphone.' Just about the same time, Thomas Edison invented the phonograph using a tinfoil-based cylinder, which was subsequently adapted to wax, and developed the Ediphone to compete directly with Columbia. The purpose of these products was to record dictation of notes and letters for a secretary (likely in a large pool that offered the service, as shown in Figure 1) who would later type them out (offline), thereby circumventing the need for costly stenographers. This turn-of-the-century concept of office mechanization spawned a range of electric and electronic implements and improvements, including the electric typewriter, which changed the face of office automation in the middle years of the twentieth century. It does not take much imagination to envision the obvious interest in creating an automatic typewriter that could directly respond to and transcribe a human's voice without having to deal with the annoyance of recording and handling the speech on wax cylinders or other recording media.

A similar kind of automation took place a century later in the 1990s in the area of call centers. A call center is a concentration of agents or associates that handle telephone calls from customers requesting assistance. Among the tasks of such call centers are routing the incoming calls to the proper department, where specific help is provided or where transactions are carried out. One example of such a service was the AT&T Operator line, which helped a caller place calls, arrange payment methods, and conduct credit card transactions. The number of agent positions (or stations) in a large call center could reach several thousand. Automatic speech recognition technologies provided the capability of automating these call handling functions, thereby reducing the large operating cost of a call center. By way of example, the AT&T Voice Recognition Call Processing (VRCP) service, which was introduced into the AT&T Network in 1992, routinely handles about 1.2 billion voice transactions with machines each year, using automatic speech recognition technology to appropriately route and handle the calls (Wilpon and Roe, 1992).

Speech recognition technology has also been a topic of great interest to a broad general population since it became popularized in several blockbuster movies of the 1960s and 1970s. In Stanley Kubrick's acclaimed movie *2001: A Space Odyssey*, an intelligent computer named HAL spoke in a natural-sounding voice and was able to recognize and understand fluently



Figure 1 An early 20th-century transcribing pool at Sears, Roebuck and Co. The women are using cylinder dictation machines, and listening to the recordings with ear-tubes. Copyright 1998–2003 by David Morton.

spoken speech, and respond as if it were an intelligent being. This anthropomorphism of HAL made the general public aware of the potential of intelligent machines. In the famous *Star Wars* saga, George Lucas extended the abilities of intelligent machines by making them mobile as well as intelligent and the droids such as R2D2 and C3PO were able to speak naturally, recognize and understand fluent speech, and move around and interact with their environment, with other droids, and with the human population at large. More recently (in 1988), in the technology community, Apple Computer created a vision of speech technology and computers for the year 2011, called Knowledge Navigator, which defined the concepts of a Speech User Interface (SUI) and a Multimodal User Interface (MUI) along with the theme of intelligent voice-enabled agents. This video had a dramatic effect in the technical community and focused technology efforts, especially in the area of visual talking agents.

Today speech technologies are commercially available for a limited but interesting range of tasks. These technologies enable machines to respond correctly and reliably to human voices, and provide useful and valuable services. While we are still far from having a machine that converses with humans on arbitrary topics like another human, many important scientific and technological advances have taken place, bringing us closer to the Holy Grail of machines that recognize and understand fluently spoken speech. This article attempts to provide a historic perspective on key inventions that have enabled progress in speech recognition and language understanding and briefly reviews several technology milestones as well as enumerating some of the remaining challenges that lie ahead of us.

From Speech Production Models to Spectral Representations

Attempts to develop machines to mimic a human's speech communication capability appear to have started in the second half of the 18th century. The early interest was not on recognizing and understanding speech, but instead on creating a speaking machine, perhaps due to the readily available knowledge of acoustic resonance tubes that were used to approximate the human vocal tract. In 1773 the Russian scientist Christian Kratzenstein, a professor of physiology in Copenhagen, succeeded in producing vowel sounds using resonance tubes connected to organ pipes (Kratzenstein, 1782: 358–380). In 1791 Wolfgang von Kempelen in Vienna constructed an Acoustic-Mechanical Speech Machine (Dudley and Tarnoczy, 1950) and in the mid-1800s Charles Wheatstone (Wheatstone, 1879) built a version of von Kempelen's speaking machine using resonators made of leather, the configuration of which could be altered or controlled with a hand to produce different speech-like sounds, as shown in Figure 2.

During the first half of the 20th century, work by Fletcher (Fletcher, 1922) and others at Bell Laboratories documented the relationship between a given speech spectrum (which is the distribution of power of a speech sound across frequency), and its sound characteristics as well as its intelligibility to a human listener. In the 1930s Homer Dudley, influenced greatly by Fletcher's research, developed a speech synthesizer called the VODER (Voice Operating Demonstrator) (Dudley and Watkins, 1939), which was an electrical equivalent (with mechanical control) of Wheatstone's mechanical speaking machine. Figure 3 shows a block diagram of Dudley's VODER,

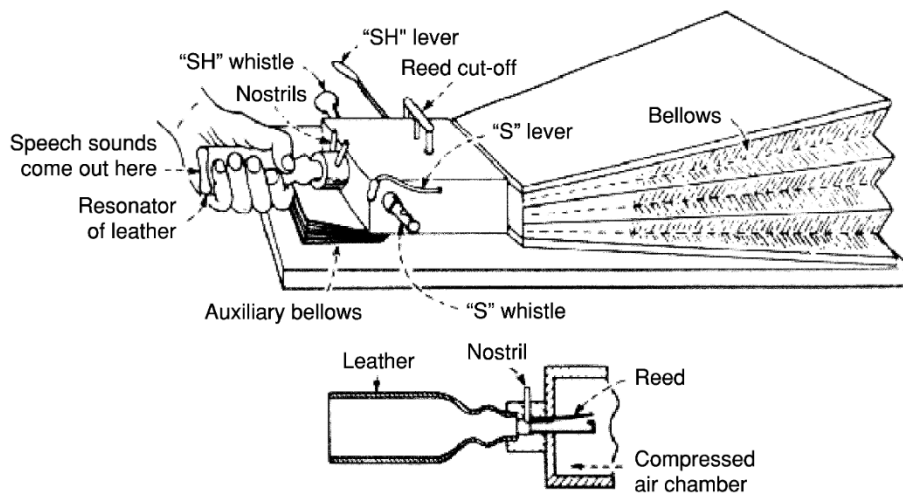


Figure 2 Wheatstone's version of von Kempelen's speaking machine (from Flanagan, 1972).

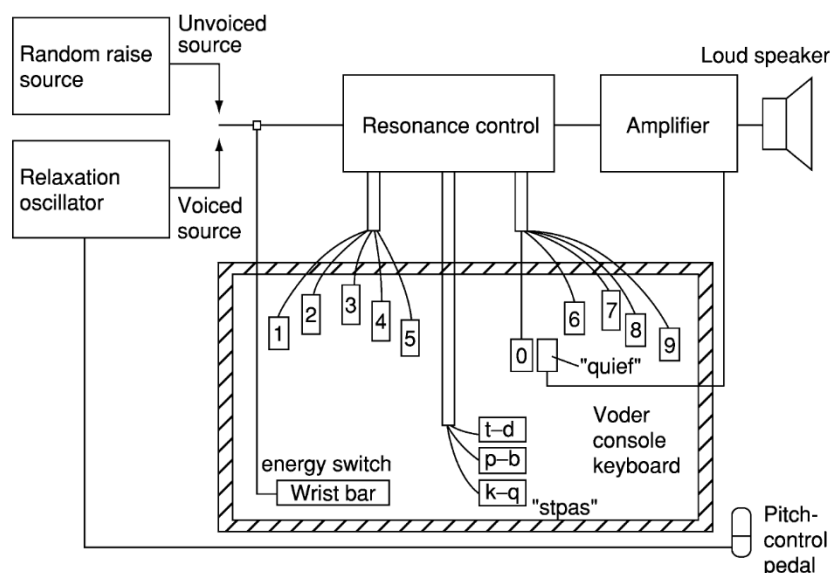


Figure 3 A block schematic of Homer Dudley's VODER (From Dudley *et al.*, 1939).



Figure 4 The VODER at the 1939 World Fair in New York City.

which consisted of a wrist bar for selecting either a relaxation oscillator output or noise as the driving signal, and a foot pedal to control the oscillator frequency (the pitch of the synthesized voice). The driving signal was passed through ten bandpass filters whose output levels were controlled by the operator's fingers. These ten bandpass filters were used to alter

the power distribution of the source signal across a frequency range, thereby determining the characteristics of the speechlike sound at the loudspeaker. Thus to synthesize a sentence, the VODER operator had to learn how to control and 'play' the VODER so that the appropriate sounds of the sentence were produced. The VODER was demonstrated at the World's Fair in New York City in 1939 (shown in [Figure 4](#)) and was considered an important milestone in the evolution of speaking machines.

The pioneering work of Harvey Fletcher and Homer Dudley firmly established the importance of the signal spectrum for reliable identification of the phonetic nature of a speech sound. Following the convention established by these two outstanding scientists, most modern systems and algorithms for speech recognition are based on the concept of measurement of the (time-varying) speech power spectrum (or its variants such as the cepstrum), in part due to the fact that measurement of the power spectrum from a signal is relatively easy to accomplish with modern digital signal processing techniques.

Early Automatic Speech Recognizers

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the phonetic elements of speech (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a

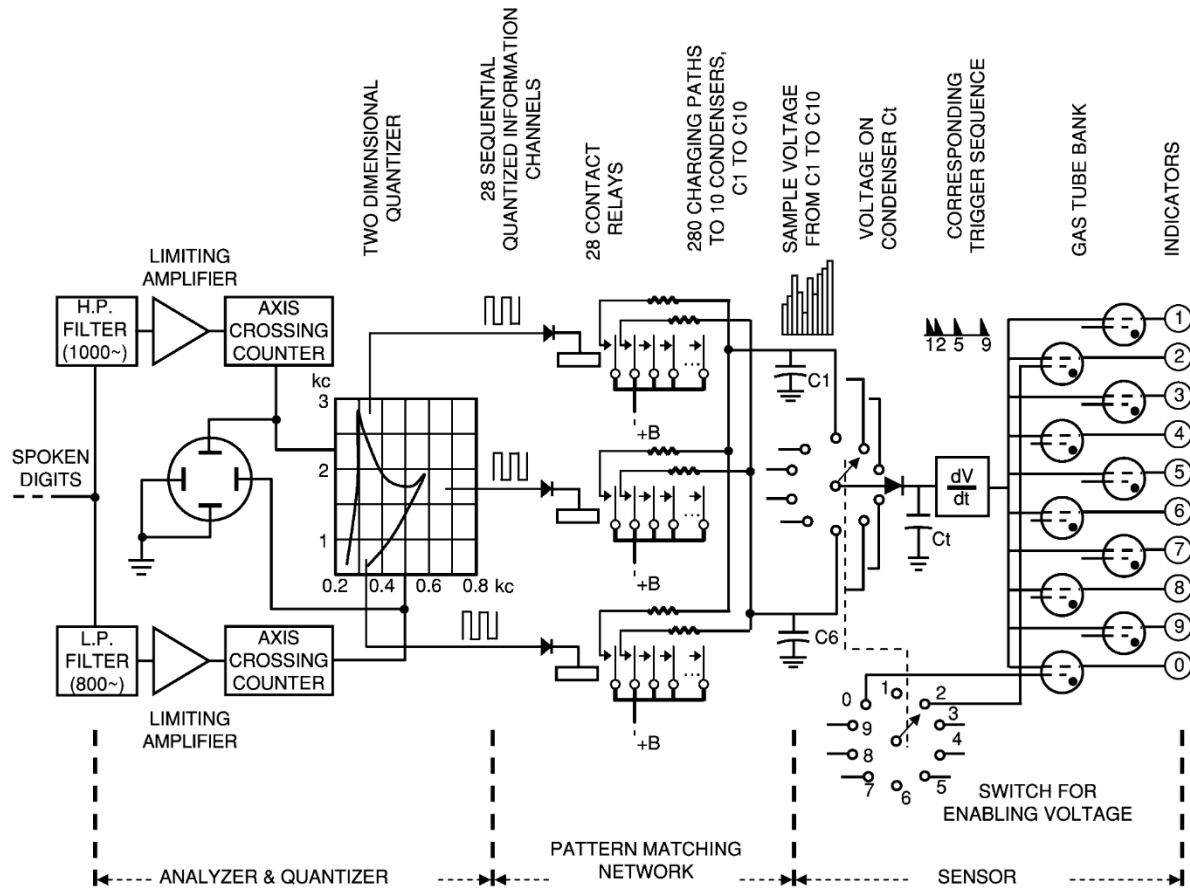


Figure 5 Block schematic of digit recognizer circuits (from Davis *et al.*, 1952).

steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube. These natural modes of resonance, called the *formants* or *formant frequencies*, are manifested as major regions of energy concentration in the speech power spectrum. In 1952 Davis *et al.*, of Bell Laboratories built a system for isolated digit recognition for a single speaker (1952) using the formant frequencies measured (or estimated) during vowel regions of each digit. Figure 5 shows a block diagram of the digit recognizer developed by Davis *et al.*, and Figure 6 shows plots of the formant trajectories along the dimensions of the first and the second formant frequencies for each of the ten digits, 1–9 and 0, respectively. These trajectories, in which time was an implicit parameter, served as the reference pattern for determining the identity of an unknown digit utterance as the best matching digit.

In other early recognition systems of the 1950s, Olson and Belar of RCA Laboratories built a system to recognize 10 syllables of a single talker (Olson and Belar, 1956) and at MIT Lincoln Lab, Forgie and

Forgie built a speaker-independent 10-vowel recognizer (Forgie and Forgie, 1959). In the 1960s, several Japanese laboratories demonstrated their capability of building special-purpose hardware to perform a speech recognition task. Most notable were the vowel recognizer of Suzuki and Nakata at the Radio Research Lab in Tokyo (Suzuki and Nakata, 1961), the phoneme recognizer of Sakai and Doshita at Kyoto University (Sakai and Doshita, 1962), and the digit recognizer of NEC Laboratories (Nagata *et al.*, 1963). The work of Sakai and Doshita involved the first use of a speech segmenter for analysis and recognition of speech in different portions of the input utterance. In contrast, an isolated digit recognizer implicitly assumed that the unknown utterance contained a complete digit (and no other speech sounds or words) and thus did not need an explicit segmenter. Kyoto University's work could be considered a precursor to a continuous speech recognition system.

In another early recognition system Fry and Denes, at University College in England, built a phoneme recognizer to recognize 4 vowels and 9 consonants (Fry and Denes, 1959). By incorporating statistical information about allowable phoneme sequences in

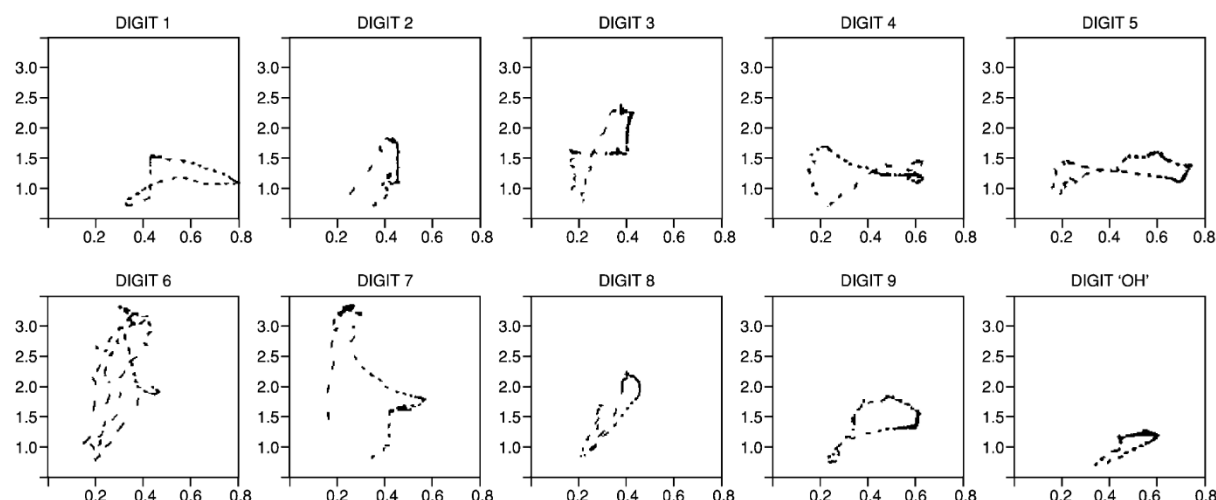


Figure 6 Photographs of formant 1 vs. formant 2 presentation of the digits. (from Davis *et al.*, 1952).

English, they increased the overall phoneme recognition accuracy for words consisting of two or more phonemes. This work marked the first use of *statistical syntax* (at the phoneme level) in automatic speech recognition.

An alternative to the use of a speech segmenter was the concept of adopting a *non-uniform time scale* for aligning speech patterns. This concept started to gain acceptance in the 1960s through the work of Tom Martin at RCA Laboratories (Martin *et al.*, 1964) and Vintsyuk in the Soviet Union (Vintsyuk, 1968). Martin recognized the need to deal with the temporal nonuniformity in repeated speech events and suggested a range of solutions, including detection of utterance endpoints, which greatly enhanced the reliability of the recognizer performance (Martin *et al.*, 1964). Vintsyuk proposed the use of dynamic programming for time alignment between two utterances in order to derive a meaningful assessment of their similarity (Vintsyuk, 1968). His work, though largely unknown in the West, appears to have preceded that of Sakoe and Chiba (1978) as well as others who proposed more formal methods, generally known as dynamic time warping, in speech pattern matching. Since the late 1970s, mainly due to the publication by Sakoe and Chiba, dynamic programming, in numerous variant forms (including the Viterbi algorithm, Viterbi 1967, which came from the communication theory community), has become an indispensable technique in automatic speech recognition.

Technology Drivers since the 1970s

In the late 1960s, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC) (Atal and Hanauer, 1971;

Itakura and Saito, 1970), which greatly simplified the estimation of the vocal tract response from speech waveforms. By the mid-1970s, the basic ideas of applying fundamental pattern recognition technology to speech recognition, based on LPC spectral estimation methods, were proposed by Itakura (1975), Rabiner and Levinson (1979), and others.

Also during this time period, based on his earlier success at aligning speech utterances, Tom Martin founded Threshold Technology, the first company to attempt to commercialize speech recognition technology, and developed the first real ASR product called the VIP-100 System. The system was only used in a few simple applications, such as by television face-plate manufacturing firms (for quality control) and by FedEx (for package sorting on a conveyor belt), but its main importance was the way it influenced the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense to fund the Speech Understanding Research (SUR) program during the early 1970s. Among the systems built by the contractors of the ARPA program was Carnegie Mellon University's (CMU) Harpy (Lowerre, 1986), which was shown to be able to recognize speech using a vocabulary of 1011 words with reasonable accuracy. One particular contribution from the Harpy system was the concept of doing a graph search, where the speech recognition language is represented as a connected graph derived from lexical representations of words, with syntactical production rules and word boundary rules. In the Harpy system, the input speech was first analyzed to produce a time-dependent parametric representation and then segmented into sounds. The segmented parametric representation of speech was then subjected to phone template matching using the Itakura distance (Itakura, 1975). The

graph search, based on a beam search algorithm, compiled, hypothesized, pruned, and then verified the recognized sequence of words (or sounds) that satisfied the knowledge constraints with the highest matching score (smallest distance to the reference patterns). The Harpy system was perhaps the first to take advantage of a finite state network to reduce computation and efficiently determine the closest matching string. However, methods that optimized the finite state network (FSN) (for performance as well as to eliminate redundancy) did not come about until the early 1990s (Mohri, 1997).

Other systems developed under DARPA's SUR program included CMU's Hearsay(-II) and BBN's HWIM (Klatt, 1977). Neither Hearsay-II nor HWIM (Hear What I Mean) met the DARPA program's performance goal at its conclusion in 1976. However, the approach proposed by Hearsay-II of using parallel asynchronous processes that simulate the component knowledge sources in a speech system was a pioneering concept. The Hearsay-II system extended sound identity analysis (to higher level hypotheses) given the detection of a certain type of (lower level) information or evidence, which was provided to a global 'blackboard' where knowledge from parallel sources was integrated to produce the next level of hypothesis. BBN's HWIM system, on the other hand, was known for introducing interesting new ideas such as a lexical decoding network incorporating sophisticated phonological rules (aimed at phoneme recognition accuracy), its handling of segmentation ambiguity by a lattice of alternative hypotheses, and the concept of word verification at the parametric level. Another system worth noting of the time was the DRAGON system by Jim Baker, who moved to Massachusetts to start a company with the same name in the early 1980s.

In parallel to the ARPA-initiated efforts, two broad directions in speech recognition research started to take shape in the 1970s, with IBM and AT&T Bell Laboratories essentially representing two different schools of thought as to the applicability of automatic speech recognition systems for commercial applications.

IBM's effort, led by Fred Jelinek, was aimed at creating a voice-activated typewriter (VAT), the main function of which was to convert a spoken sentence into a sequence of letters and words that could be shown on a display or typed on paper (Jelinek *et al.*, 1975). This type of speech recognition task is generally referred to as *transcription*. The recognition system, called Tangora, was essentially a speaker-dependent system (i.e., the typewriter had to be trained by each user). The technical focus was on the size of the recognition vocabulary (with dictation in office

correspondence as the primary application), and the structure of the language (the grammar). IBM addressed the issue of grammar from a statistical decoding perspective; statistical syntactical rules were used to describe how likely, in a probabilistic sense, was a sequence of language symbols (e.g., phonemes or words) that could appear in the speech signal. The set of statistical grammatical or syntactical rules was called a *language model*, of which the *n*-gram model, which defined the probability of occurrence of an ordered sequence of *n* words, was the most frequently used variant. Although both the *n*-gram language model and a traditional grammar are manifestations of the rules of the language, their roles were fundamentally different. The *n*-gram model, which characterized the word relationship within a span of *n* words, was purely a convenient and powerful statistical representation of a grammar. Its effectiveness in guiding a word search for speech recognition, however, was strongly validated by the famous word game of Claude Shannon (1948) that involved a competition between a human and a computer. In this competition both the computer and the human are asked to sequentially guess the next word in an arbitrary sentence. The human guesses based on native experience with language; the computer uses the accumulated word statistics to make its best guess based on maximum probability from the estimated word frequencies. It was shown that once the span of the words, *n*, exceeded three, the computer was very likely to win (make better guesses as to the next word in the sequence) over the human player. Since their introduction in the 1980s, the use of *n*-gram language models, and their variants, has become indispensable in large vocabulary speech recognition systems.

At AT&T Bell Laboratories, the goal of the research program was to provide automated telecommunication services to the public, such as voice dialing and command and control for routing of phone calls. These automated systems were expected to work well for a vast population (tens of millions) of talkers. The focus at Bell Laboratories was in the design of a speaker-independent system that could deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with notably different regional accents. This led to the creation of a range of speech clustering algorithms for creating word and sound reference patterns (initially templates but ultimately statistical models) that could be used across a wide range of talkers and accents. Furthermore, research to understand and to control the acoustic variability of various speech representations across talkers led to the study of a range of spectral distance measures (e.g., the Itakura distance [Itakura, 1975] and statistical modeling

techniques [Juang *et al.* 1986]) that produced sufficiently rich representations of the utterances from a vast population. (As will be elaborated in the next section, the technique of mixture density hidden Markov models [Juang, 1985; Lee *et al.*, 1990] has since become the prevalent representation of speech units for speaker independent continuous speech recognition.) Since applications such as voice dialing and call routing usually involved only short utterances of limited vocabulary and consisted of only a few words, there was an emphasis of the research at Bell Laboratories on what is generally called the *acoustic model* (the spectral representation of sounds or words) over the language model (the representation of the grammar or syntax of the task). Also, of great importance in the Bell Laboratories' approach was the concept of *keyword spotting* as a primitive form of speech understanding (Wilpon, 1990). The technique of keyword spotting aimed at detecting a keyword or a key phrase of some particular significance that was embedded in a longer utterance where there was no semantic significance to the other words in the utterance. The need for such keyword spotting was to accommodate talkers who preferred to speak in natural sentences rather than using rigid command sequences when requesting services (i.e., as if they were speaking to a human operator). For example, a telephone caller requesting a credit card charge might speak the sentence 'I'd like to charge it to my credit card' rather than just say 'credit card.' In a limited domain application, the presence of the key-phrase 'credit card' in an otherwise naturally spoken sentence was generally sufficient to indicate the caller's intent to make a credit card call. The detected keyword or key-phrase would then trigger a prescribed action (or sequence of actions) as part of the service, in response to the talker's spoken utterance. The technique of keyword spotting required extension of the usual pattern recognition paradigm to one that supported hypothesis testing.

The IBM and AT&T Bell Laboratories approaches to speech recognition both had a profound influence in the evolution of human-machine speech communication technology of the last two decades. One common theme between these efforts, despite the differences, was that mathematical formalism and rigor started to emerge as distinct and important aspects of speech recognition research. While the difference in goals led to different realizations of the technology in various applications, the rapid development of statistical methods in the 1980s, most notably the hidden Markov model (HMM) framework (Jelinek, 1976; Levinson *et al.*, 1983), caused a certain degree of convergence in the system design. Today, most practical speech recognition systems are based on the

statistical framework and results developed in the 1980s, with significant additional improvements in the 1990s.

Technology Directions in the 1980s and 1990s

Speech recognition research in the 1980s was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) toward a more rigorous statistical modeling framework. Although the basic idea of the hidden Markov model (HMM) was known and understood early on in a few laboratories (e.g., IBM and the Institute for Defense Analyses (IDA) [Ferguson, 1980]), the methodology was not complete until the mid-1980s and it wasn't until after widespread publication of the theory (Levinson *et al.*, 1983; Ferguson, 1980) that the hidden Markov model became the preferred method for speech recognition. The popularity and use of the HMM as the main foundation for automatic speech recognition and understanding systems has remained constant over the past two decades, especially because of the steady stream of improvements and refinements of the technology.

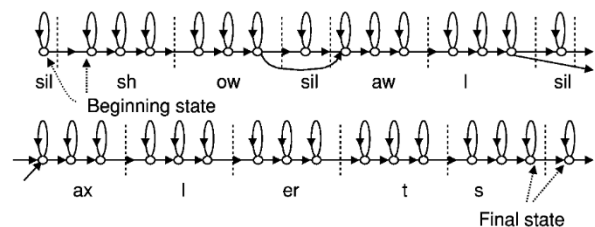
The hidden Markov model, which is a doubly stochastic process, models the intrinsic variability of the speech signal (and the resulting spectral features) as well as the structure of spoken language in an integrated and consistent statistical modeling framework (Rabiner and Juang, 2004). As is well known, a speech signal is inherently highly variable (due to variations in pronunciation and accent, as well as environmental factors such as reverberation and noise). When people speak the same word, the acoustic signals are not identical (in fact they may even be remarkably different at the acoustic level), even though the underlying linguistic structure, in terms of the pronunciation, syntax and grammar, may (or may not) remain the same. The formalism of the HMM is a probability measure that uses a Markov chain to represent the linguistic structure and a set of probability distributions to account for the variability in the acoustic realization of the sounds in the utterance. Given a set of known (text-labeled) utterances, representing a sufficient collection of the variations of the words of interest (called a training set), one can use an efficient estimation method, called the Baum-Welch algorithm (Baum, 1972), to obtain the 'best' set of parameters that define the corresponding model or models. The estimation of the parameters that define the model is equivalent to training and learning. The resulting model is then used to provide an indication of the likelihood (probability) that an

unknown utterance is indeed a realization of the word (or words) represented by the model. The probability measure represented by the hidden Markov model is an essential component of a speech recognition system that follows the statistical pattern recognition approach, and has its root in Bayes's decision theory (Theodoridis and Koutroumbas, 2003). The HMM methodology represented a major step forward from the simple pattern recognition and acoustic-phonetic methods used earlier in automatic speech recognition systems.

The idea of the hidden Markov model appears to have originated in the late 1960s at the Institute for Defense Analyses (IDA) in Princeton, N.J. Len Baum referred to an HMM as a set of probabilistic functions of a Markov chain, which, by definition, involves two nested distributions, one pertaining to the Markov chain and the other to a set of the probability distributions, each associated with a state of the Markov chain, respectively (Baum, 1972). The HMM model attempts to address the characteristics of a probabilistic sequence of observations that may not be a fixed function but instead changes according to a Markov chain. This doubly stochastic process was found to be useful in a number of applications such as stock market prediction and crypto-analysis of a rotary cipher (an encryption device widely used during World War II). Baum's modeling and estimation technique was first shown to work for discrete observations (i.e., ones that assume values from a finite set and thus are governed by discrete probability distributions) and then for random observations that were well modeled using log-concave probability density functions. The technique was powerful but limited. Using a representation theorem by Fan (Liporace, 1982), Liporace, also of IDA, relaxed the log-concave density constraint to include an elliptical symmetric density constraint (thereby including a Gaussian density and a Cauchy density). Baum's doubly stochastic process started to find applications in the speech area, initially in speaker identification systems, in the late 1970s (Poritz, 1982; Liporace, 1982). As more people attempted to use the HMM technique, it became clear that the constraint on the form of the density functions imposed a limitation on the performance of the system, particularly for speaker independent tasks where the speech parameter distribution was not sufficiently well modeled by a simple log-concave or an elliptically symmetric density function. In the early 1980s at Bell Laboratories, the theory of HMM was extended to mixture densities (Juang, 1985; Juang *et al.*, 1986) that have since proven vitally important in ensuring satisfactory recognition accuracy, particularly for speaker independent, large vocabulary speech recognition tasks.

The HMM, being a probability measure, was amenable for incorporation in a larger speech decoding framework which included a language model. The use of a finite-state grammar in large vocabulary continuous speech recognition represented a consistent extension of the Markov chain that the HMM utilized to account for the structure of the language, albeit at a level that accounted for the interaction between articulation and pronunciation. Although these structures (for various levels of the language constraints) were at best crude approximations to the real speech phenomenon, they were computationally efficient and often sufficient to yield reasonable (first-order) performance results. The merger of the hidden Markov model (with its advantage in statistical consistency, particularly in handling acoustic variability) and the finite state network (with its search and computational efficiency, particularly in handling word sequence hypotheses) was an important, although not unexpected, technological development in the mid-1980s.

Figure 7 shows a finite state composite model for the utterance 'show all alerts,' constructed from several context-dependent subword models that represent the corresponding phoneme-like speech units (including a unit for silence [sil] that can occur at the beginning and end of the sentence, as well as at the end of any word in the sentence, as might occur during a pause in speaking). The finite state graph is realized as a Markov chain for calculation of the likelihood, based on the observation sequence (the spectral representation over time) of an unknown utterance. Note that each node in the graph is associated with a probability distribution which accounts for the variability in realizing the corresponding phoneme-like sound. The likelihood that an utterance was generated by the finite state network represented by the model is computed as a sequential sum of local likelihoods related to the elementary units of the composite model. This sequential sum is obtained as a result of optimal state alignment, via a dynamic programming procedure, to maximize the match between the labeled units and the corresponding



"Show all alerts" modeled as phones: ϕ -sh-ow, ϕ -ax-l, ax-l-er, l-er-t

Figure 7 A composite finite-state network for the utterance 'show all alerts.'

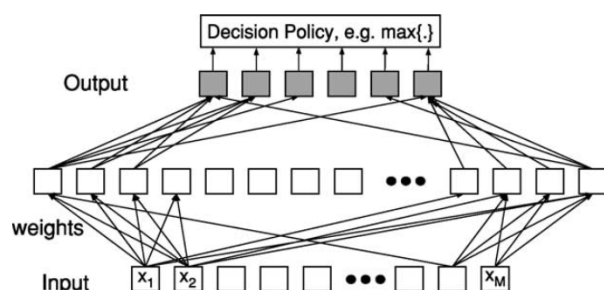


Figure 8 Multi-layer perceptron.

portions of the speech observations. At any given time, there are a number of hypothesized units whose likelihoods are to be evaluated and the determination of sound identity is based on the maximum likelihood value (or score of the match). The number of hypothesized units for match and that of the paths for search can be at times astronomical and thus may require efficient computational algorithms to solve the problem. A tool called FSM (finite-state machine library), which embodies the above finite state network in a unified transducer framework for incorporation of weighted search was developed in the mid-1990s (Mohri, 1997).

Another technology that was reintroduced in the late 1980s was the idea of artificial neural networks (ANN). Neural networks were first introduced in the 1950s, but failed to produce notable results initially (McCulloch and Pitts, 1943). The advent in the 1980s of a parallel distributed processing (PDP) model, which was a dense interconnection of simple computational elements, and a corresponding 'training' method, called error back-propagation, revived interest in the old idea of mimicking the human neural processing mechanism. A particular form of PDP, the multi-layer perceptron, shown in Figure 8, received perhaps the most intense attention then, not because of its analog to neural processing but due to its capability in approximating any function (of the input) to an arbitrary precision, provided no limitation in the complexity of the processing configuration was imposed. If a pattern recognizer is viewed as one that performs a function that maps an input pattern to its class identity, the multi-layer perceptron was then a readily available candidate for this purpose. Early attempts at using neural networks for speech recognition centered on simple tasks like recognizing a few phonemes or a few words (e.g., isolated digits), with good success (Lippmann, 1990). However, as the problem of speech recognition inevitably requires handling of temporal variation, neural networks in their original form have not proven to be extensible to this task. Ongoing research focuses on integrating neural networks with the essential

structure of a hidden Markov model to take advantage of its temporal handling capability.

In the 1990s a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error (Juang *et al.*, 1997). This fundamental change of paradigm was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory would become inapplicable under these circumstances. After all, the objective of a recognizer design should be to achieve the least recognition error rather than the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. The concept of minimum classification or empirical error subsequently spawned a number of techniques, among which discriminative training and kernel-based methods such as the support vector machines (SVM) have become popular subjects of study (Juang *et al.*, 1997; Bahl *et al.*, 1986; Vapnik, 1998) and have found applications in speech recognition.

The success of statistical methods revived the interest from (D)ARPA at the juncture of the 1980s and the 1990s, leading to several new speech recognition systems including the Sphinx system from CMU (Lee, 1988), the BYBLOS system from BBN (Schwartz *et al.*, 1989) and the DECIPHER system from Stanford Research Institute (SRI) (Murveit *et al.*, 1989). CMU's Sphinx system successfully integrated the statistical method of hidden Markov models with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context-dependent phone models in a sophisticated lexical decoding network, achieving remarkable results for large-vocabulary continuous speech recognition.

With the support of DARPA, evaluation of speech recognition technology for a wide range of tasks and task vocabularies was diligently pursued throughout the 1990s and into the twenty-first century. Such evaluations were mostly based on the measurement of word (and sentence) error rate as the performance figure of merit of the recognition system. Furthermore, these evaluations were conducted systematically over carefully designed tasks with progressive degrees of difficulty, ranging from the recognition of continuous speech spoken with stylized grammatical structure (as used routinely in military tasks, for example, the Naval Resource Management task); to transcriptions of live (off-the-air) news broadcast, for example, North American Business news, which involves a fairly large vocabulary over 20 000 words;

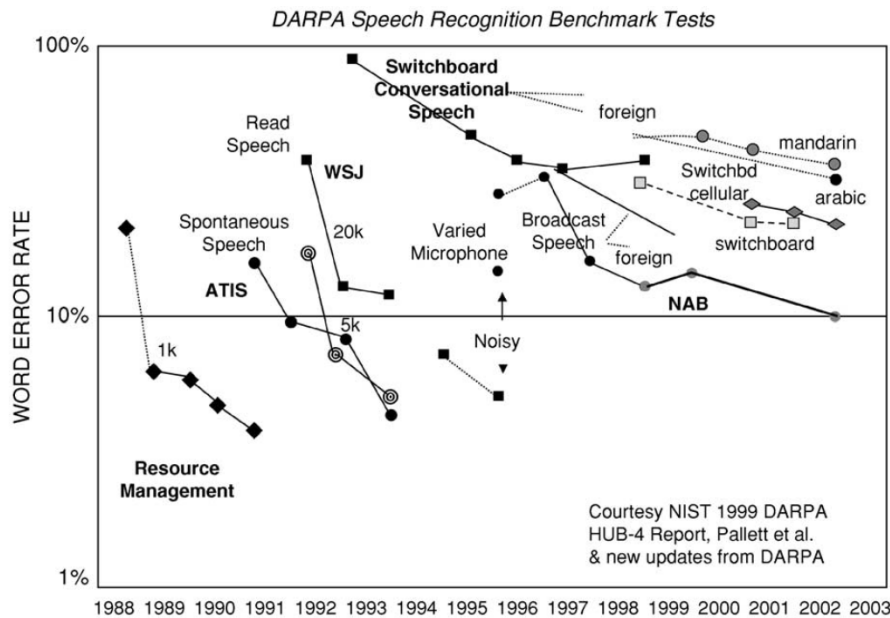


Figure 9 DARPA benchmark evaluation of speech recognition for a number of tasks.

and conversational speech. Figure 9 shows a chart that summarizes the benchmark performance of various large vocabulary continuous speech recognition tasks, as measured in formal DARPA and National Institute of Standards and Technology (NIST) evaluations (Pallett *et al.*, 1995). In the chart, the task of Resource Management involves a rigidly stylized military expression with a vocabulary of nearly 1000 words. ATIS is a task that involves simple spontaneous speech conversation with an automated air travel information retrieval system; although the speech is spontaneous, its linguistic structure is rather limited in scope. WSJ refers to transcription of a spoken (read) paragraph from the *Wall Street Journal*; the vocabulary size could be as large as 60 000 words. The Switchboard task is one of the most challenging ones proposed by DARPA. The speech is conversational and spontaneous, with many instances of the so-called disfluencies such as partial words, hesitation, and repairs. The general conclusion that can be drawn from these results is that conversational speech, which does not strictly adhere to linguistic constraints, is significantly more difficult to recognize than task-oriented speech that follows strict syntactic and semantic production rules. Also, the evaluation program showed that increasing the amount of speech data used for estimating the recognizer parameters (i.e., the size of the training set) always led to reductions of word error rate. (It is a well-accepted target that in order for virtually any large vocabulary speech recognition task to become viable, the word error rate must fall below a 10% level).

In the 1990s great progress was made in the development of software tools that enabled many individual research programs all over the world. As systems became more sophisticated (many large vocabulary systems now involve tens of thousands of phone unit models and millions of parameters), a well-structured baseline software system became indispensable for further research and development to incorporate new concepts and algorithms. The system that was made available by the Cambridge University team (led by Steve Young), called the Hidden Markov Model Tool Kit (HTK) (Young *et al.*), was one of the most widely adopted software tools for automatic speech recognition research. Other widely available tool kits for HMM modeling of speech include the Sphinx System from Carnegie-Mellon University (<http://cmusphinx.sourceforge.net/sphinx4/>), and the HMM toolkit from Mississippi State University (<http://www.isip.msstate.edu/projects/speech/>).

Toward a Machine that Communicates

Most speech recognition research, up to the 1980s, considered the major research problem to be one of converting a speech waveform (as an acoustic realization of a linguistic event) into words (as a best-decoded sequence of linguistic units). Many researchers also believed that the speech-to-text process was the necessary first step in the process that enabled a machine to be able to understand and properly respond to human speech. In field evaluations of speech recognition and understanding

technology for a range of tasks, two important things were learned about the speech communication process between humans and machines. First, potential users of a speech recognition system tended to speak natural sentences that often did not fully satisfy the grammatical constraints of the recognizer (e.g., by including out-of-vocabulary [OOV] words, non-grammatical constructs, ill-formed sentences, etc.), and the spoken utterances were also often corrupted by linguistically irrelevant noise components such as ambient noise, extraneous acoustic sounds, interfering speech, etc. Second, as in human-to-human speech communications, speech applications often required a dialog between the user and the machine to reach some desired state of understanding. Such a dialog often required such operations as query and confirmation, thus providing some allowance for speech recognition and understanding errors. The keyword spotting method (and its application in AT&T's Voice Recognition Call Processing (VRCP) System, as mentioned earlier), was introduced in response to the first factor while the second factor focused the attention of the research community on the area of dialog management. Many applications and system demonstrations that recognized the importance of dialog management over a system's raw word recognition accuracy were introduced in the early 1990s with the goal of eventually creating a machine that really mimicked the communicating capabilities of a human. Among these systems, Pegasus and Jupiter developed at the Massachusetts Institute of Technology under Victor Zue were particularly noteworthy demos (Glass and Weinstein, 2001; Zue *et al.*, 2000), and the How May I Help You? (HMIHY[®]) system at AT&T developed by Allen Gorin was an equally noteworthy service that was introduced as part of AT&T Customer Care for their Consumer Communications Services in 2000 (Gorin *et al.*, 1996).

Pegasus is a speech conversational system that provides information about the status of airline flights over an ordinary telephone line. Jupiter is a similar system with a focus on weather information access, both local and national. These systems epitomized the effectiveness of dialog management. With properly designed dialog management, these systems could guide the user to provide the required information to process a request, among a small and implicit set of menu choices, without explicitly requesting details of the query, e.g., such as by using the dialog management phrase "please say morning, afternoon, or evening" when time frame of the flight was solicited. Dialog management also often incorporated imbedded confirmation of recognized phrases and soft error handling so as to make the user react as if there was a real human agent rather than a machine

on the other end of the telephone line. The goal was to design a machine that communicated rather than merely recognized the words in a spoken utterance.

The late 1990s was marked by the deployment of real speech-enabled applications, ranging from AT&T's VRCP (automated handling of operator-assisted calls) and Universal Card Service (customer service line) that were used daily (often by millions of people) in lieu of a conventional voice response system with touch-tone input, to United Airlines' automatic flight information system by SpeechWorks and AT&T's "How May I Help You? (HMIHY)" system for call routing of consumer help line calls. Although automatic speech recognition and speech understanding systems are far from perfect in terms of the word or task accuracy, properly developed applications can still make good use of the existing technology to deliver real value to the customer, as evidenced by the number and extent of such systems that are used on a daily basis by millions of users.

Summary and Outlook

Figure 10 shows a timeline of progress in speech recognition and understanding technology over the past several decades. We see that in the 1960s we were able to recognize small vocabularies (10–100 words) of isolated words, based on simple acoustic-phonetic properties of speech sounds. The key technologies that were developed during this time frame were filter-bank analyses, simple time normalization methods, and the beginnings of sophisticated dynamic programming methodologies. In the 1970s we were able to recognize medium vocabularies (100–1000 words) using simple template-based, pattern recognition methods. The key technologies that were developed during this period were the pattern recognition models, the introduction of LPC methods for spectral representation, the pattern clustering methods for speaker-independent recognizers, and the introduction of dynamic programming methods for solving connected word recognition problems. In the 1980s we started to tackle large vocabulary (1000–unlimited number of words) speech recognition problems based on statistical methods, with a wide range of networks for handling language structures. The key technologies introduced during this period were the hidden Markov model (HMM) and the stochastic language model, which together enabled powerful new methods for handling virtually any continuous speech recognition problem efficiently and with high performance. In the 1990s we were able to build large vocabulary systems with unconstrained language models, and constrained task syntax models for continuous speech recognition and

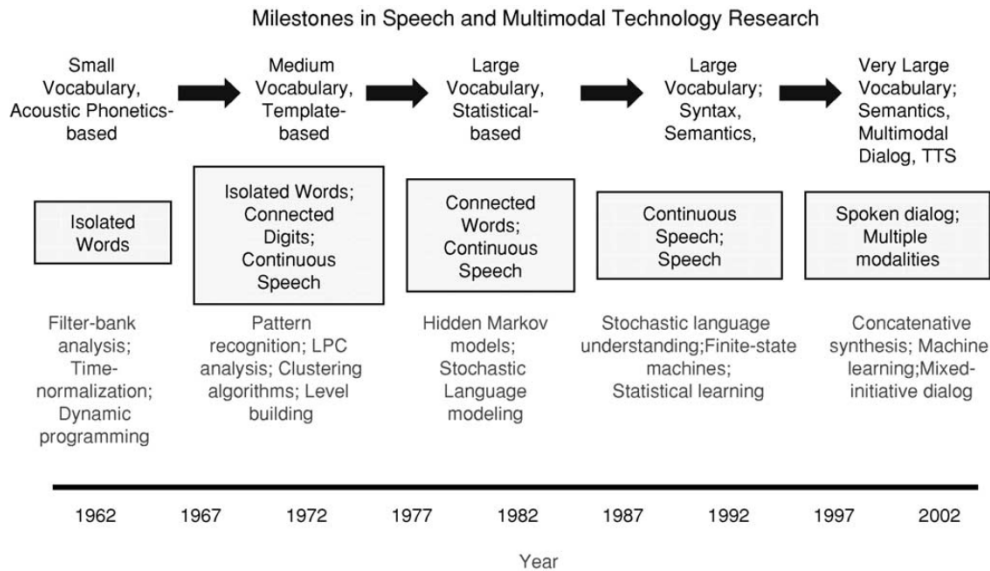


Figure 10 Milestones in speech recognition and understanding technology over the past 40 years.

understanding. The key technologies developed during this period were the methods for stochastic language understanding, statistical learning of acoustic and language models, and the introduction of finite state transducer framework (and the FSM Library) and the methods for their determination and minimization for efficient implementation of large vocabulary speech understanding systems. Finally, in the last few years, we have seen the introduction of very large vocabulary systems with full semantic models, integrated with text-to-speech (TTS) synthesis systems, and multi-modal inputs (pointing, keyboards, mice, etc.). These systems enable spoken dialog systems with a range of input and output modalities for ease-of-use and flexibility in handling adverse environments where speech might not be as suitable as other input-output modalities. During this period we have seen the emergence of highly natural concatenative speech synthesis systems, the use of machine learning to improve both speech understanding and speech dialogs, and the introduction of mixed-initiative dialog systems to enable user control when necessary.

After nearly five decades of research, speech recognition technologies have finally entered the marketplace, benefiting the users in a variety of ways. Throughout the course of development of such systems, knowledge of speech production and perception was used in establishing the technological foundation for the resulting speech recognizers. Major advances, however, were brought about in the 1960s and 1970s via the introduction of advanced speech representations based on LPC analysis and cepstral

analysis methods, and in the 1980s through the introduction of rigorous statistical methods based on hidden Markov models. All of this came about because of significant research contributions from academia, private industry, and the government. As the technology continues to mature, it is clear that many new applications will emerge and become part of our way of life—thereby taking full advantage of machines that are partially able to mimic human speech capabilities.

The challenge of designing a machine that truly functions like an intelligent human is still a major one. Our accomplishments to date are only the beginning and it will take many years before a machine can pass the Turing test, namely, achieving performance that rivals that of a human.

See also: Speech Recognition: Statistical Methods.

Bibliography

- Atal B S & Hanauer S L (1971). 'Speech analysis and synthesis by linear prediction of the speech wave.' *J. Acoust. Soc. Am.* 50(2), 637–655.
- Bahl L R, Brown P F, de Souza P V & Mercer R L (1986). 'Maximum mutual information estimation of hidden Markov model parameters for speech recognition.' *Proceedings ICASSP 86*. Tokyo. 49–52.
- Baum L E (1972). 'An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes.' *Inequalities* 3, 1–8.
- Carnegie-Mellon University, <http://cmusphinx.sourceforge.net/sphinx4/>.

- Das S K & Picheny M A (1996). 'Issues in practical large vocabulary isolated word recognition: The IBM Tangora system.' In Lee C H, Soong F K & Paliwal K K (eds.) *Automatic speech and speaker recognition advanced topics*. Boston: Kluwer. 457–479.
- Davis K H, Biddulph R & Balashek S (1952). Automatic Recognition of Spoken Digits.' *J. Acoust. Soc. Am.* 24(6), 627–642.
- Dudley H (1939). 'The Vocoder.' *Bell Labs Record* 17, 122–126.
- Dudley H, Reisz R R & Watkins S A (1939). 'A synthetic speaker.' *J. Franklin Institute* 227, 739–764.
- Dudley H & Tarnoczy T H (1950). 'The speaking machine of Wolfgang von Kempelen.' *J. Acoust. Soc. Am.* 22, 151–166.
- Ferguson J D (1980). 'Hidden Markov analysis: an introduction.' In *Hidden Markov models for speech, Institute for Defense Analyses*. Princeton, NJ.
- Flanagan J L (1972). *Speech analysis, synthesis and perception* (2nd ed.). Springer-Verlag.
- Fletcher H (1922). 'The nature of speech and its interpretations.' *Bell Syst. Tech. J.* 1, 129–144.
- Forgie J W & Forgie C D (1959). 'Results obtained from a vowel recognition computer program.' *J. Acoust. Soc. Am.* 31(11), 1480–1489.
- Fry D B & Denes P (1959). 'The design and operation of the mechanical speech recognizer at University College London.' *J. British Inst. Radio Engr.* 19(4), 211–229.
- Glass J & Weinstein E (2001). 'SpeechBuilder: facilitating spoken dialogue system development.' *7th European Conference on Speech Communication and Technology*, Aalborg, Denmark.
- Itakura F & Saito S (1970). 'A statistical method for estimation of speech spectral density and formant frequencies' *Electronics and Communications in Japan* 53A, 36–43.
- Itakura F (1975). Minimum prediction residual principle applied to speech recognition.' *IEEE Transactions on Acoustics, Speech and Signal Proceedings ASSP-23*, 57–72.
- Jelinek F (1976). 'Continuous speech recognition by statistical methods.' *Proceedings IEEE* 64, 532–536.
- Jelinek F, Bahl L R & Mercer R L (1975). Design of a linguistic statistical decoder for the recognition of continuous speech.' *IEEE Transactions on Information Theory IT-21*, 250–256.
- Juang B H, Lee C H & Chou W (1997). 'Minimum classification error rate methods for speech recognition.' *IEEE Transactions on Speech & Audio Processing, T-SA* 5(3), 257–265.
- Juang B H (1985). 'Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains.' *AT&T Tech. J.* 64(6), 1235–1249.
- Juang B H, Levinson S E & Sondhi M M (1986). 'Maximum likelihood estimation for multivariate mixture observations of Markov chains.' *IEEE Transactions on Information Theory It-32(2)*, 307–309.
- Klatt D H (1977). 'Review of the DARPA speech understanding project (1).' *J. Acoust. Soc. Am.* 62, 1345–1366.
- Kratzenstein C G (1782). 'Sur la naissance de la formation des voyelles.' *J. Phys.* 21, 358–380.
- Lee C H, Rabiner L R, Pieraccini R & Wilpon J G (1990). 'Acoustic modeling for large vocabulary speech recognition.' *Computer Speech & Language* 4, 1237–1265.
- Lee K F (1988). 'Large-vocabulary speaker-independent continuous speech recognition: the Sphinx system.' Ph.D. thesis, Carnegie Mellon University.
- Levinson S E, Rabiner L R & Sondhi M M (1983). 'An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition.' *Bell Syst. Tech. J.* 62(4), 1035–1074.
- Liporace L A (1982). Maximum likelihood estimation for multivariate observations of Markov sources.' *IEEE Transactions on Information Theory IT-28(5)*, 729–734.
- Lippmann R P (1990). 'Review of neural networks for speech recognition.' In Waibel A & Lee K F (eds.) *Readings in speech recognition*. Morgan Kaufmann Publishers. 374–392.
- Lowerre B (1986). 'The HARP Y speech understanding system.' In Lea W (ed.) *Trends in speech recognition*. Speech Science Publications. [Reprinted in Waibel A, Lee K F (eds.) *Readings in speech recognition*. Morgan Kaufmann Publishers, 1990. 576–586.]
- Martin T, Nelson A I & Zadell H J (1964). Speech recognition by feature abstraction techniques.' *Tech. Report AL-TDR-64-176*. Air Force Avionics Lab.
- McCullough W S & Pitts W H (1943). A logical calculus of ideas immanent in nervous activity.' *Bull. Math Biophysics* 5, 115–133.
- Mohri M (1997). 'Finite-state transducers in language and speech processing.' *Computational Linguistics* 23(2), 269–312.
- Murveit H, Cohen M, Price P, Baldwin G, Weintraub M & Bernstein J (1989). 'SRI's DECIPHER System.' In *Proceedings of the Speech and Natural Language Workshop*. Philadelphia. 238–242.
- Nagata K, Kato Y & Chiba S (1963). 'Spoken Digit Recognizer for Japanese Language.' *NEC Res. Develop.* 6.
- Olson H F & Belar H (1956). 'Phonetic Typewriter.' *J. Acoust. Soc. Am.* 28(6), 1072–1081.
- Pallett D S et al. (1994–1995). 'Benchmark tests for the ARPA spoken language program.' *Proceedings of the 1995 ARPA Human Language Technology Workshop*, 5–36.
- Gorin A L, Parker B A, Sachs R M & Wilpon J G (1996). 'How may I help you?' *Proceedings of Interactive Voice Technology for Telecommunications Applications (IVTTA)*, 57–60.
- Piccone J. Mississippi State University. <http://www.isip.mssstate.edu/projects/speech/>.
- Poritz A B (1982). Linear predictive hidden Markov models and the speech signal.' In *Proceedings ICASSP-82*. 1291–1294.

- Rabiner L R & Juang B H (2006). 'Statistical methods for the recognition and understanding of speech.' *Encyclopedia of language and linguistics*.
- Rabiner L R, Levinson S E, Rosenberg A E & Wilpon J G (1979). 'Speaker independent recognition of isolated words using clustering techniques.' *IEEE Transactions on Acoustics, Speech and Signal Proceedings ASSP-27*, 336–349.
- Sakai J & Doshita S (1962). 'The phonetic typewriter.' *Information Processing Proceedings IFIP Congress*, Munich.
- Sakoe H & Chiba S (1978). 'Dynamic programming algorithm quantization for spoken word recognition, IEEE Trans.' *IEEE Transactions on Acoustics, Speech and Signal Proceedings ASSP-26(1)*, 43–49.
- Schwartz R, Barry C, Chow Y-L, Derr A, Feng M-W, Kimball O, Kubala F, Makhoul J & Vandegrift J (1989). 'The BBN BYBLOS continuous speech recognition system.' In *Proceedings of the Speech and Natural Language Workshop*. Philadelphia. 94–99.
- Shannon C (1948). 'A mathematical theory of communication.' *Bell System Technical Journal* 27, 379–423; 623–656.
- Suzuki J & Nakata K (1961). 'Recognition of Japanese vowels – preliminary to the recognition of speech.' *J. Radio Res. Lab* 37(8), 193–212.
- Theodoridis S & Koutroumbas K (2003). *Pattern recognition* (2nd edn.). Elsevier Academic Press.
- Vapnik V N (1998). *Statistical learning theory*. John Wiley and Sons.
- Vintsyuk T K (1968). 'Speech discrimination by dynamic programming.' *Kibernetika* 4(2), 81–88.
- Viterbi A J (1967). 'Error bounds for convolutional codes and an asymptotically optimal decoding algorithm.' *IEEE Transactions on Information Theory IT-13*, 260–269.
- Wheatstone C (1879). *The scientific papers of Sir Charles Wheatstone*. London: Taylor and Francis.
- Wilpon J G & Roe D B (1992). 'AT&T Telephone network applications of speech recognition. *Proceedings COST232 Workshop*, Rome, Italy, Nov.
- Wilpon J G, Rabiner L R, Lee C H & Goldman E R (1990). 'Automatic recognition of keywords in unconstrained speech using hidden Markov models.' *IEEE Transactions on Acoustics, Speech and Signal Processing* 38(11), 1870–1878.
- Young S *et al.*, the HTKBook, <http://htk.eng.cam.ac.uk/>.
- Zue V *et al.* (2000). 'Jupiter: a telephone-based conversational interface for weather information, IEEE Trans.' *IEEE Transactions on Speech and Audio Processing* X, 100–112.

Speech Recognition: Psychology Approaches

S L Mattys, University of Bristol, Bristol, UK

© 2006 Elsevier Ltd. All rights reserved.

What could be simpler than perceiving and recognizing speech? Without any real effort, we identify half a dozen spoken words per second, almost immediately understand the message generated by the sentence they form, and often elaborate appropriate verbal and nonverbal responses before the utterance ends. However, empirical investigation reveals that the perceptual and cognitive mechanisms involved in recognizing speech are far from obvious. The computational complexity that listeners face is experienced first hand by second-language learners and by speech scientists trying to build speech-recognition machines. The problem originates mainly from the nature of the speech signal itself.

Main Properties of the Speech Signal

The fact that speech is conveyed in the auditory modality has dramatic implications for the psychological operations underpinning its recognition. Auditory language processing and written language

processing are subjected to such different constraints that the models put forward to account for one or the other overlap only minimally. Over the past 50 years, research in speech recognition has focused on the constraining effects of three main properties of the auditory signal: sequentiality, variability, and continuity.

Sequentiality

One of the most obvious disadvantages of the auditory system compared to the visual system is that the distribution of the auditory information is solely under the speaker's control. From the listener's perspective, the speech signal is transient, and there is no such thing as auditory regressions, in which listeners can replay the words or sentences they just heard. The timing constraints on how speech sounds reach our ears have direct consequences for the time course of speech processing. Because we cannot hold more than a few syllables or words at a time in our auditory short-term memory, word and sentence processing must take place sequentially, in real time. For a large part, listeners are extremely good at keeping up with the rapid flow of speech sounds: Marslen-Wilson (1984) showed that words in sentences are often