Automatic Speech Recognition – A Brief History of the Technology Development

*B.H. Juang[#] & Lawrence R. Rabiner**
[#] Georgia Institute of Technology, Atlanta
** Rutgers University and the University of California, Santa Barbara*

Abstract

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. Since the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis [1, 2], the problem of automatic speech recognition has been approached progressively, from a simple machine that responds to a small set of sounds to a sophisticated system that responds to fluently spoken natural language and takes into account the varying statistics of the language in which the speech is produced. Based on major advances in statistical modeling of speech in the 1980s, automatic speech recognition systems today find widespread application in tasks that require a human-machine interface, such as automatic call processing in the telephone network and query-based information systems that do things like provide updated travel information, stock price quotations, weather reports, etc. In this article, we review some major highlights in the research and development of automatic speech recognition during the last few decades so as to provide a technological perspective and an appreciation of the fundamental progress that has been made in this important area of information and communication technology.

Keywords

Speech recognition, speech understanding, statistical modeling, spectral analysis, hidden Markov models, acoustic modeling, language modeling, finite state network, office automation, automatic transcription, keyword spotting, dialog systems, neural networks, pattern recognition, time normalization

1. 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 one hundred 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 "*Edi*phone" 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 mid-part 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 1990's 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 in-coming 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 [3].

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 (David Morton, the history of Sound Recording History, http://www.recording-history.org/)

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 1960's and 1970's, most notably Stanley Kubrick's acclaimed movie "2001: A Space Odyssey". In this movie, an intelligent computer named "HAL" spoke in a natural sounding voice and was able to recognize and understand fluently spoken speech, and respond accordingly. 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 like 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, titled "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 any topic 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 an 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.

2. 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 2nd 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 which 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 [4]. Later, Wolfgang von Kempelen in Vienna constructed an "Acoustic-Mechanical Speech Machine" (1791) [5] and in the mid-1800's Charles Wheatstone [6] 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.

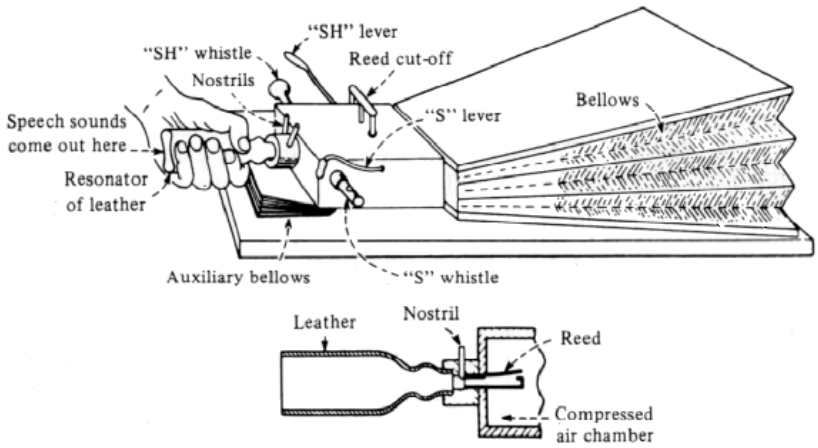

Figure 2 Wheatstone's version of von Kempelen's speaking machine (Flanagan [7]).

During the first half of the 20th century, work by Fletcher [8] 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, as perceived by a human listener. In the 1930's Homer Dudley, influenced greatly by Fletcher's research, developed a speech synthesizer called the VODER (Voice Operating Demonstrator) [2], which was an electrical equivalent (with mechanical control) of Wheatstone's mechanical speaking machine. Figure 3 shows a block diagram of Dudley's VODER 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 speech-like 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 Fair in New York City in 1939 (shown in Fig 4) and was considered an important milestone in the evolution of speaking machines.

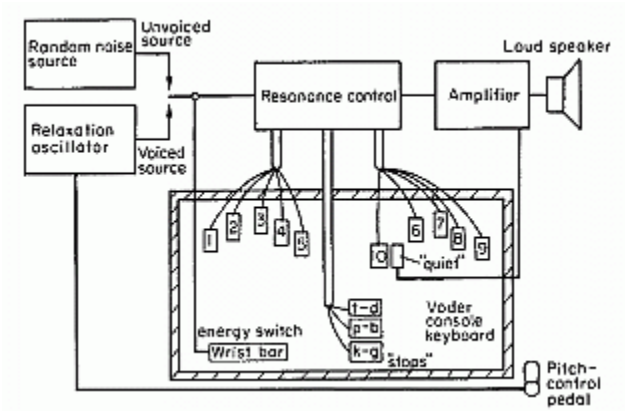

Figure 3 A block schematic of Homer Dudley's VODER [2].

Speech pioneers like Harvery 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.

Figure 4 The VODER at the 1939 World's Fair in NYC.

3. 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 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, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition for a single speaker [9], 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, one-nine and oh, respectively. These trajectories served as the "reference pattern" for determining the identity of an unknown digit utterance as the best matching digit.

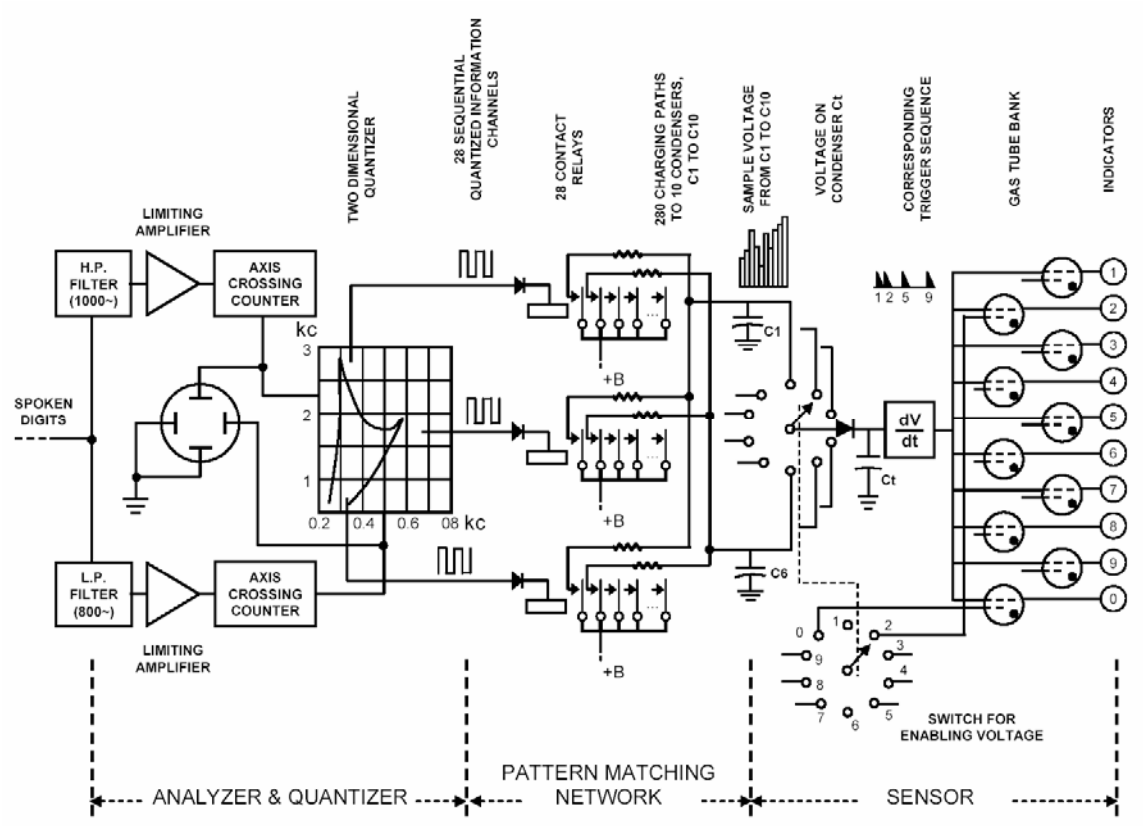

Figure 5 Block schematic of digit recognizer circuits. (Davis, Biddulph, and Balashek [9])

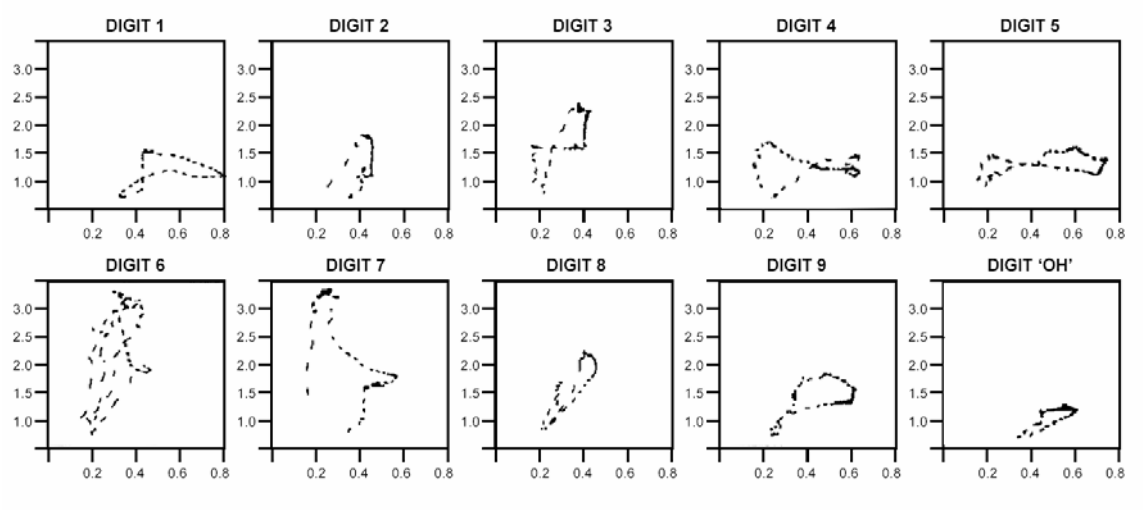

Figure 6 Photographs of formant 1 vs. formant 2 presentation of the digits. (Davis, Biddulph, and Balashek [9])

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

Forgie built a speaker-independent 10-vowel recognizer [11]. In the 1960's, 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 [12], the phoneme recognizer of Sakai and Doshita at Kyoto University [13], and the digit recognizer of NEC Laboratories [14]. 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 [15]. By incorporating statistical information about allowable phoneme sequences in 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 1960's through the work of Tom Martin at RCA Laboratories [16] and Vintsyuk in the Soviet Union [17]. Martin recognized the need to deal with the temporal non-uniformity in repeated speech events and suggested a range of solutions, including detection of utterance endpoints, which greatly enhanced the reliability of the recognizer performance [16]. Vintsyuk proposed the use of dynamic programming for time alignment between two utterances in order to derive a meaningful assessment of their similarity [17]. His work, though largely unknown in the West, appears to have preceded that of Sakoe and Chiba [18] as well as others who proposed more formal methods, generally known as dynamic time warping, in speech pattern matching. Since the late 1970's, mainly due to the publication by Sakoe and Chiba, dynamic programming, in numerous variant forms (including the Viterbi algorithm [19] which came from the communication theory community), has become an indispensable technique in automatic speech recognition.

4. Technology Drivers since the 1970's

In the late 1960's, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC) [20, 21], which greatly simplified the estimation of the vocal tract response from speech waveforms. By the mid 1970's, the basic ideas of applying

fundamental pattern recognition technology to speech recognition, based on LPC methods, were proposed by Itakura [22], Rabiner and Levinson [23] and others.

Also during this time period, based on his earlier success at aligning speech utterances, Tom Martin founded the first speech recognition commercial company called Threshold Technology, Inc. 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 faceplate 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 1970's. Among the systems built by the contractors of the ARPA program was Carnegie Mellon University's "Harpy" (Lowerre [24]) which was shown to be able to recognize speech using a vocabulary of 1,011 words, and with reasonable accuracy. One particular contribution from the Harpy system was the concept of doing a graph search, where the speech recognition language was represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. In the proposed Harpy system, the input speech, after going through a parametric analysis, was segmented and the segmented parametric sequence of speech was then subjected to phone template matching using the Itakura distance [22]. 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 which optimized the resulting finite state network (FSN) (for performance as well as to eliminate redundancy) did not come about until the early 1990's [25] (see section 5).

Other systems developed under DARPA's SUR program included CMU's Hearsay(-II) and BBN's HWIM [26]. 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 its interesting ideas including 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 1970's, 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 [27]. The recognition system, called Tangora, was essentially a speaker-dependent system (i.e., the typewriter had to be trained by each individual user). The technical focus was on the size of the recognition vocabulary (as large as possible, with a primary target being one used in office correspondence), and the structure of the language model (the grammar), which was represented by statistical syntactical rules that described how likely, in a probabilistic sense, was a sequence of language symbols (e.g., phonemes or words) that could appear in the speech signal. This type of speech recognition task is generally referred to as *transcription*. 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 [28] which 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 3, 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 1980's, the use of *n*-gram language models, and its 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 (literally tens of millions) of talkers without the need for individual speaker training. 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 [22]) and statistical modeling techniques [30] that produced sufficiently rich representations of the utterances from a vast population. (As will be discussed in the next section, the technique of mixture density hidden Markov models [31, 32] 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 [33]. 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 1980's, most notably the hidden Markov model (HMM) framework [34-35], 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 1980's, with significant additional improvements in the 1990's.

5. Technology Directions in the 1980's and 1990's

Speech recognition research in the 1980's was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) towards 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) [36]), the methodology was not complete until the mid-1980's and it wasn't until after widespread publication of the theory [35-36] 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 [37]. As is well known, a realistic 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), 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 [38], 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' decision theory [39]. 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 first come out in the late 1960's 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 [38]. 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, which was 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 random observations that were well modeled using log-concave probability density functions. The technique was powerful but limited. 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), with help from an old representation theorem by Fan [41]. Baum's doubly stochastic process started to find applications in the speech area, initially in speaker identification systems, in the late 1970's [40-41]. 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 1980's at Bell Laboratories, the theory of HMM was extended to mixture densities [30-31] which 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-1980's.

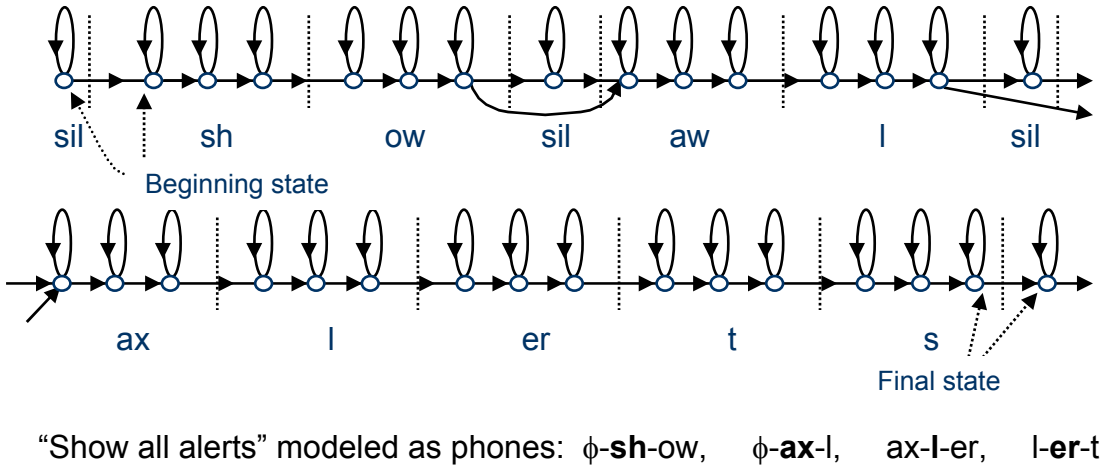

Figure 7 A composite finite-state network for the utterance "show all alerts."

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 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 elementary units of the composite model) after a dynamic programming state alignment is performed to maximize the match between the labeled units and the corresponding portions of the speech observations (even for models of incorrect word sequences), respectively. At any given time, there are a number of hypothesized units 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 the FSM (finite-state machine) library, which embodied the finite state network approach in a unified transducer framework (including weighted search) was developed in the mid-1990s [25] and has been a major component of almost all modern speech recognition and understanding systems.

Another technology that was (re)introduced in the late 1980's was the idea of artificial neural networks (ANN). Neural networks were first introduced in the 1950's, but failed to produce notable results initially [42]. The advent, in the 1980's, 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 around the old idea of mimicking the human neural processing mechanism. A particular form of PDP, the multi-layer perceptron, shown in Fig. 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 mapping 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 [43]. 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. On-going research focuses on integrating neural networks with the essential structure of a hidden Markov model to take advantage of the temporal handling capability of the HMM.

In the 1990's, 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 [44]. 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 [44-46].

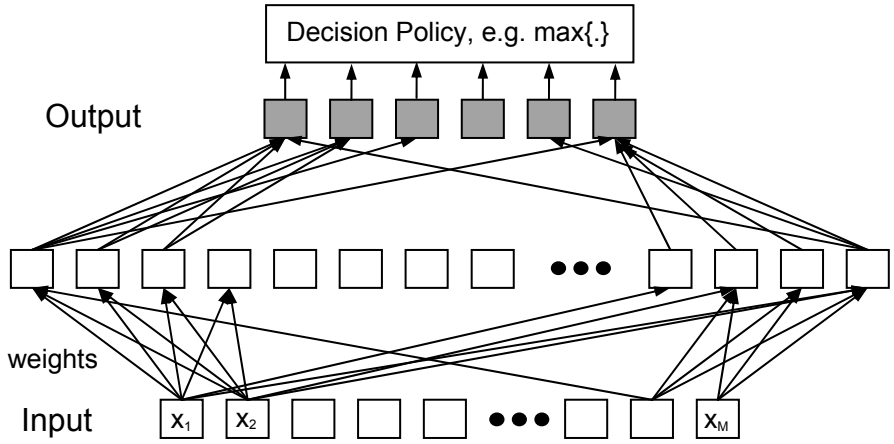

Figure 8 Multi-layer Perceptron

The success of statistical methods revived the interest from DARPA at the juncture of the 1980's and the 1990's, leading to several new speech recognition systems including the Sphinx system from CMU [47], the BYBLOS system from BBN [48] and the DECIPHER system from SRI [49]. 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 1990's 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, e.g., the Naval Resource Management task) to transcriptions of live (off-the-air) news broadcast (e.g., NAB that involves a fairly large vocabulary over 20K words) 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 NIST evaluations [50]. 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 set of spoken (read) paragraphs from the Wall Street Journal; the vocabulary size could be as large as 60K 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, etc. 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 1990's 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 was 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) [51], was (and remains today as) one of the most widely adopted software tools for automatic speech recognition research.

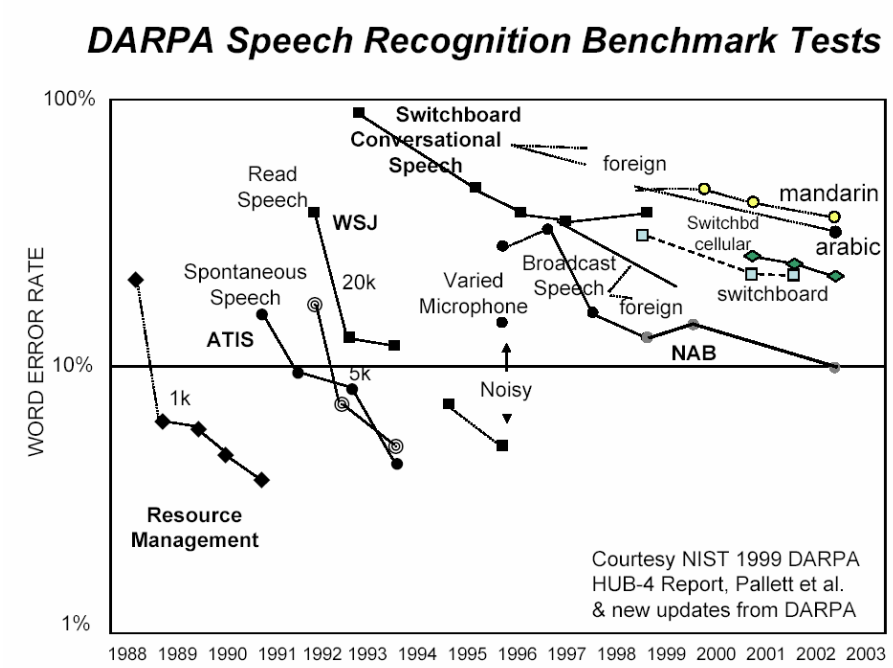

Figure 9 DARPA benchmark evaluation of speech recognition for a number of tasks.

6. Towards a Machine That Communicates

Most speech recognition research, up to the 1980's, 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 1990's 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 [52,53], and the How May I Help You (HMIHY) system at AT&T developed by Al Gorin was an equally noteworthy service that was introduced as part of AT&T Customer Care for their Consumer Communications Services in 2000 [54].

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 1990's 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 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.

7. Summary & Outlook

Figure 10 shows a timeline of progress in speech recognition and understanding technology over the past several decades. We see that in the 1960's we were able to recognize small

vocabularies (order of 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 1970's we were able to recognize medium vocabularies (order of 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 1980's 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 1990's we were able to build large vocabulary systems with unconstrained language models, and constrained task syntax models for continuous speech recognition and 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 1960's and 1970's via the introduction of advanced speech representations based on LPC analysis and cepstral analysis methods, and in the 1980's 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 going forward. 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.

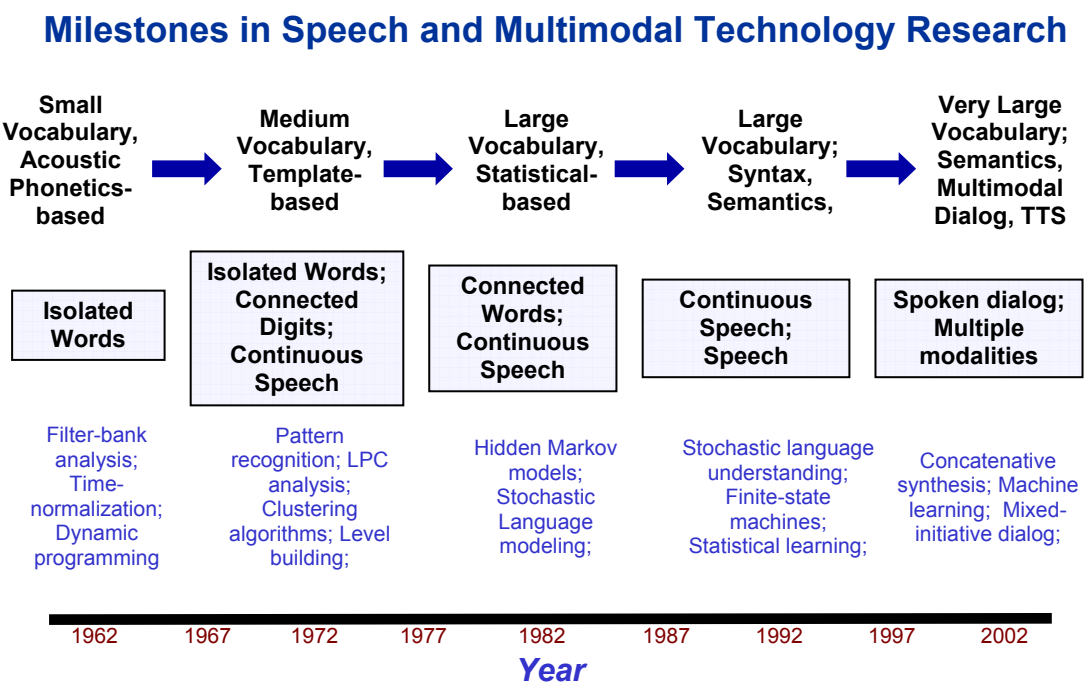

Figure 10 Milestones in Speech Recognition and Understanding Technology over the Past 40 Years.

References

1. H. Dudley, *The Vocoder*, Bell Labs Record, Vol. 17, pp. 122-126, 1939.
2. H. Dudley, R. R. Riesz, and S. A. Watkins, *A Synthetic Speaker*, J. Franklin Institute, Vol. 227, pp. 739-764, 1939.
3. J. G. Wilpon and D. B. Roe, *AT&T Telephone Network Applications of Speech Recognition*, Proc. COST232 Workshop, Rome, Italy, Nov. 1992.

4. C. G. Kratzenstein, *Sur la raissance de la formation des voyelles*, J. Phys., Vol 21, pp. 358-380, 1782.

5. H. Dudley and T. H. Tarnoczy, *The Speaking Machine of Wolfgang von Kempelen*, J. Acoust. Soc. Am., Vol. 22, pp. 151-166, 1950.

6. Sir Charles Wheatstone, *The Scientific Papers of Sir Charles Wheatstone*, London: Taylor and Francis, 1879.

7. J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Second Edition, Springer-Verlag, 1972.

8. H. Fletcher, *The Nature of Speech and its Interpretations*, Bell Syst. Tech. J., Vol 1, pp. 129-144, July 1922.

9. K. H. Davis, R. Biddulph, and S. Balashek, *Automatic Recognition of Spoken Digits*, J. Acoust. Soc. Am., Vol 24, No. 6, pp. 627-642, 1952.

10. H. F. Olson and H. Belar, *Phonetic Typewriter,* J. Acoust. Soc. Am., Vol. 28, No. 6, pp. 1072-1081, 1956.

11. J. W. Forgie and C. D. Forgie, *Results Obtained from a Vowel Recognition Computer Program*, J. Acoust. Soc. Am., Vol. 31, No. 11, pp. 1480-1489, 1959.

12. J. Suzuki and K. Nakata, *Recognition of Japanese Vowels—Preliminary to the Recognition of Speech*, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212, 1961.

13. J. Sakai and S. Doshita, *The Phonetic Typewriter,* Information Processing 1962, Proc. IFIP Congress, Munich, 1962.

14. K. Nagata, Y. Kato, and S. Chiba, *Spoken Digit Recognizer for Japanese Language*, NEC Res. Develop., No. 6, 1963.

15. D. B. Fry and P. Denes, *The Design and Operation of the Mechanical Speech Recognizer at University College London*, J. British Inst. Radio Engr., Vol. 19, No. 4, pp. 211-229, 1959.

16. T. B. Martin, A. L. Nelson, and H. J. Zadell, *Speech Recognition by Feature Abstraction Techniques*, Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

17. T. K. Vintsyuk, *Speech Discrimination by Dynamic Programming*, Kibernetika, Vol. 4, No. 2, pp. 81-88, Jan.-Feb. 1968.

18. H. Sakoe and S. Chiba, *Dynamic Programming Algorithm Quantization for Spoken Word Recognition*, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.

19. A. J. Viterbi, *Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm*, IEEE Trans. Informaiton Theory, Vol. IT-13, pp. 260-269, April 1967.

20. B. S. Atal and S. L. Hanauer, *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*, J. Acoust. Soc. Am. Vol. 50, No. 2, pp. 637-655, Aug. 1971.

21. F. Itakura and S. Saito, *A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies,* Electronics and Communications in Japan, Vol. 53A, pp. 36-43, 1970.

22. F. Itakura, *Minimum Prediction Residual Principle Applied to Speech Recognition*, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-23, pp. 57-72, Feb. 1975.

23. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon, *Speaker Independent Recognition of Isolated Words Using Clustering Techniques*, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. Assp-27, pp. 336-349, Aug. 1979.

24. B. Lowerre, *The HARPY Speech Understanding System*, Trends in Speech Recognition, W. Lea, Editor, Speech Science Publications, 1986, reprinted in Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, pp. 576-586, Morgan Kaufmann Publishers, 1990.

25. M. Mohri, *Finite-State Transducers in Language and Speech Processing,* Computational Linguistics, Vol. 23, No. 2, pp. 269-312, 1997.

26. Dennis H. Klatt, *Review of the DARPA Speech Understanding Project (1)*, J. Acoust. Soc. Am., 62, 1345-1366, 1977.

27. F. Jelinek, L. R. Bahl, and R. L. Mercer, *Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech*, IEEE Trans. On Information Theory, Vol. IT-21, pp. 250-256, 1975.

28. C. Shannon, *A mathematical theory of communication*, Bell System Technical Journal*,* vol. 27, pp. 379-423 and 623-656, July and October, 1948.

29. S. K. Das and M. A. Picheny, Issues in practical large vocabulary isolated word recognition: The IBM Tangora system, in *Automatic Speech and Speaker Recognition Advanced Topics*, C.H. Lee, F. K. Soong, and K. K. Paliwal, editors, p. 457-479, Kluwer, Boston, 1996.

30. B. H. Juang, S. E. Levinson and M. M. Sondhi, *Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains,* IEEE Trans. Information Theory, Vol. It-32, No. 2, pp. 307-309, March 1986.

31. B. H. Juang, *Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains*, AT&T Tech. J., Vol. 64, No. 6, pp. 1235-1249, July-Aug. 1985.

32. C.H. Lee, L.R, Rabiner, R. Pieraccini, and J.G. Wilpon, *Acoustic modeling for large vocabulary speech recognition*, Computer Speech & Language, 4: 1237-1265, January 1990.

33. J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, *Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models*, IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. 38, No. 11, pp. 1870-1878, November 1990.

34. F. Jelinek, *Continuous Speech Recognition by Statistical Methods*, Proc. IEEE, Vol. 64, pp. 532-536, April 1976.

35. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, *An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition*, Bell Syst. Tech. J., Vol. 62, No. 4, pp. 1035-1074, April 1983.

36. J. D. Ferguson, *Hidden Markov Analysis: An Introduction*, in Hidden Markov Models for Speech, Institute for Defense Analyses, Princeton, NJ 1980.

37. L. R. Rabiner and B. H. Juang, *Statistical Methods for the Recognition and Understanding of Speech,* Encyclopedia of Language and Linguistics, 2004.

38. L. E. Baum, *An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes*, Inequalities, Vol. 3, pp. 1-8, 1972.

39. S. Theodoridis and K. Koutroumbas, *Pattern Recognition: Second Edition,* Elsevier Academic Press, 2003.

40. A.B. Poritz, *Linear predictive hidden Markov models and the speech signal*, in Proc. ICASSP-82, 1291-1294, Paris France, 1982.

41. L. A. Liporace, *Maximum Likelihood Estimation for Multivariate Observations of Markov Sources*, IEEE Trans. On Information Theory, Vol. IT-28, No. 5, pp. 729-734, 1982.

42. W. S. McCullough and W. H. Pitts, *A Logical Calculus of Ideas Immanent in Nervous Activity,* Bull. Math Biophysics, Vol. 5, pp. 115-133, 1943.

43. R. P. Lippmann, *Review of Neural Networks for Speech Recognition*, Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, pp. 374-392, 1990

44. B.H. Juang, C.H. Lee and Wu Chou, *Minimum classification error rate methods for speech recognition,* IEEE Trans. Speech & Audio Processing, T-SA, vo.5, No.3, pp.257-265, May 1997.

45. L. R. Bahl, P. F. Brown, P. V. deSouza and L. R. Mercer, *Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition*, Proc. ICASSP 86, Tokyo, Japan, pp. 49-52, April 1986.

46. V. N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

47. K.-F. Lee, *Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system*, Ph.D. Thesis, Carnegie Mellon University, 1988.

48. R. Schwartz and C. Barry and Y.-L. Chow and A. Derr and M.-W. Feng and O. Kimball and F. Kubala and J. Makhoul and J. Vandegrift, *The BBN BYBLOS Continuous Speech Recognition System,* in Proc. of the Speech and Natural Language Workshop, p. 94-99, Philadelphia, PA, 1989.

49. H. Murveit and M. Cohen and P. Price and G. Baldwin and M. Weintraub and J. Bernstein, *SRI's DECIPHER System*, in proceedings of the Speech and Natural Language Workshop, p.238-242, Philadelphia, PA, 1989.

50. D. S. Pallett et al, *1994 Benchmark Tests for the ARPA Spoken Language Program,* Proc. Of the 1995 ARPA Human Language Technology Workshop, pp. 5-36, 1995.

51. S. Young, et. al., the *HTKBook*, http://htk.eng.cam.ac.uk/.

52. J. Glass and E. Weinstein, *SpeechBuilder: Facilitating Spoken Dialogue System Development,* 7th European Conf. on Speech Communication and Technology, Aalborg Denmark, Sept. 2001.

53. V. Zue et al, *Jupiter: A Telephone-Based Conversational Interface for Weather Information*, IEEE Trans. On Speech and Audio Processing, Vol. X, pp. 100-112, Jan. 2000.

54. A. L. Gorin, B. A. Parker, R. M. Sachs, and J. G. Wilpon, *How May I Help You?*, Proc. Interactive Voice Technology for Telecommunications Applications (IVTTA), pp. 57-60, Oct. 1996.