The Past, Present, and Future of Speech Processing

The Speech Processing Technical Committee

Edited by
B.H. Juang

The special series for the 50th anniversary of the Signal Processing Society continues in this issue with an article that covers the domain of the Speech Processing Technical Committee. This article provides a succinct review of the history and current status of the field of speech-processing research and describes future contributions speech processing will make to society.

Because speech is the most natural form of human communication, speech processing has been one of the most exciting areas of signal processing. In the last several decades, speech research has drawn scientists and engineers together to form an important discipline. It has created many technical impacts on society. Speech-coding algorithms have made voice communication and the storage of voice data effective and efficient. Speech-recognition technology has made it possible for computers to follow human voice commands and even understand human languages. Speech-synthesis techniques have created many interactive systems that correspond with humans with a natural voice. As computers become faster and more ubiquitous, these and other areas in speech processing are expected to flourish further and bring about an era of true human-computer interaction.

To summarize the exciting developments in this field, the article presents an insightful review and reports the authors' views in the various areas of speech processing. Topics covered in this article include speech analysis and synthesis, speech coding, speech enhancement, speech recognition, spoken-language understanding, speaker identification and verification, and multimodal communication. In addition, a sidebar reviews the history of secure voice coding.

I invite you to read this article to review the history of speech processing, to understand its current trends, and to foresee its future prospects envisioned by experts in the field. Enjoy!

Tsuhan Chen, Guest Editor
Carnegie Mellon University

1053-5888/98/$10.00©1998IEEE

As part of the celebration for the 50th anniversary of the IEEE Signal Processing Society, this article intends to provide a succinct review of speech research, in particular its history, current trends, and prospects for the future. The research areas covered are speech analysis and synthesis, speech coding, speech enhancement, speech recognition, spoken language understanding, speaker identification and verification, and multimodal communication. We omit from this discussion such topics as speech perception and production and related physiological aspects, not because they are not a part of speech research, but in order to bound the scope of the effort and to cover those topics most related to readers of this magazine. We hope readers of *IEEE Signal Processing Magazine* as well as members of the IEEE Signal Processing Society will be able to draw a picture of this important area of research and to appreciate its significance, particularly from the signal-processing perspective. We must caution the reader that such a review is cursory at best and may suffer from errors of judgement and omission.

This article was commissioned by the Speech Technical Committee of the Signal Processing Society. Many renowned speech-communication researchers were invited to contribute to this article. The list of authors represents those who submitted written contributions.

Speech Analysis and Synthesis

Research in speech processing and communication, for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. The earliest attempt of this type was a mechanical mimic of the human vocal apparatus by Wolfgang von Kempelen, described in his book published in 1791 [1]. Charles Wheatstone, some 40 years later, constructed a machine based on Kempelen's specification using a bellows to represent the lung in providing a reservoir of compressed air [2]. The vocal cords were replaced by a vibrating reed that was placed at one end of a flexible leather tube—the "vocal tract"—whose cross-sectional area could be varied to produce various voiced sounds. Other sounds could be produced by the machine as well, e.g., nasals by opening a side branch tube (the "nostrils"), fricatives by shutting off the reed and introducing turbulence at appropriate places in the vocal tract, and stops by closing the tube and opening it abruptly. It appears that Wheatstone was able to produce a fairly large repertoire of vowels and consonants and even some short sentences using this simple mechanical device.

Contributing Authors

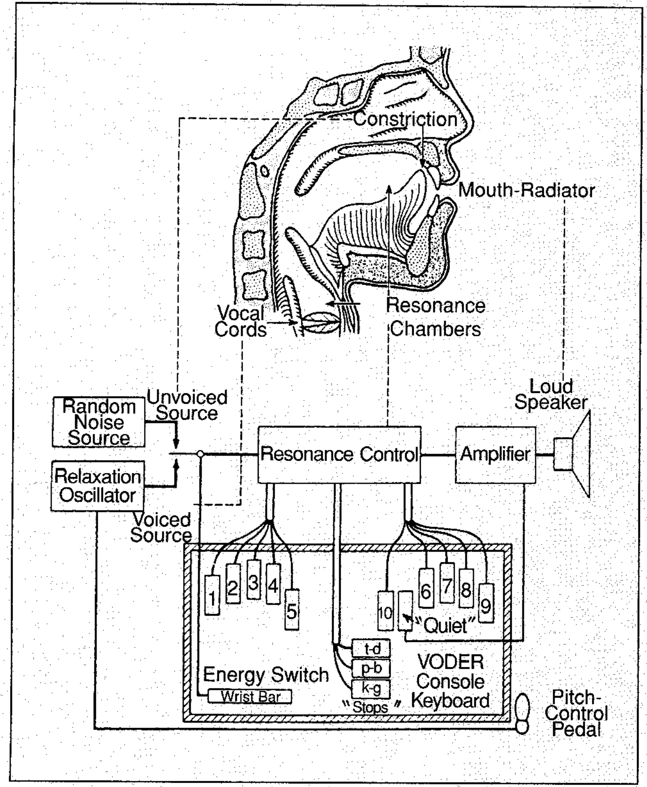
Don Childers, *University of Florida, Gainesville, USA*
R.V. Cox, *AT&T Labs-Research, USA*
Renato DeMori, *University of Avignon, France*
Sadaoki Furui, *Tokyo Institute of Technologies, Japan*
B.H. Juang, *Bell Labs, Lucent Technologies, USA*
J.J. Mariani, *LIMSI, France*
Patti Price, *SRI, USA*
Shigeki Sagayama, *NTT, Japan*
M.M. Sondhi, Bell Labs, *Lucent Technologies, USA*
Ralph Weischedel, *BBN/GTE, USA*

Interest in mechanical analogs of the human vocal apparatus continued into the 20th century. While several notable people (Faber, Bell, Paget, and Riesz) followed Kempelen and Wheatstone's speech-production models, Helmholz, Miller, Koenig, and others pursued a different design principle. They synthesized vowel sounds by superimposing harmonically related sinusoids with appropriately adjusted amplitudes. These two fundamentally different approaches, source-tract modeling (motivated by physics) and sinusoidal modeling (motivated by mathematics), have dominated the speech signal-processing field for more than 100 years.

Research interest in speech processing today has gone well beyond the simple notion of mimicking the human vocal apparatus (which still intrigues many researchers). The scope (both breadth and depth) of speech research today has become much larger due to advances in mathematical tools (algorithms), computers, and the almost limitless potential applications of speech processing in modern communication systems and networking. Conversely, speech research has been viewed as an important driving force behind many of the advances in computing and software engineering, including digital signal processors (DSPs). Such a synergetic relationship will continue for years to come.

Source-Tract and Source-Filter Modeling

Source-tract modeling by electrical circuits, realized in the form of a source-filter system, was first proposed by Homer Dudley at Bell Laboratories in the 1930s [3]. As an electrical engineer, Dudley exploited his insights in modulated-carrier radio transmission to construct an *electrical* speech synthesizer that dispensed with all the mechanical devices of von Kempelen's synthesizer. A highly simplified, but accurate, schematic of Dudley's synthesizer is shown in Fig. 1. The electrical excitation source had two components—a "buzz" source (for voiced speech) and a "hiss" source (for unvoiced speech). The buzz source was a relaxation oscillator that generated a sequence of pulses with a controllable repetition rate (the fundamental frequency) and provided the voiced carrier. The hiss source was the shot noise generated by a vacuum tube, and it provided the unvoiced carrier. The message (i.e., the time-varying characteristics of the vocal tract) was modulated on the source carrier by passing the output of the source through a filter whose frequency response was adjustable. This variable filter was realized by a bank (10 channels) of bandpass filters covering the range of speech frequencies. Any desired vocal-tract



▲ *1. Schematic diagram of the VODER synthesizer (after Dudley, Riesz, Watkins and Flanagan [2]).*

frequency-response characteristic was achieved by adjusting the amplitudes of the outputs of the bandpass filters.

With the collaboration of Riesz and Watkins, Dudley implemented two highly acclaimed devices, the VODER (VOice DEmonstration Recorder) and the VOCODER, based on this principle. The VODER (a schematic diagram of which is shown in Fig.1) was a system in which an operator manipulated a keyboard with 14 keys, a wrist bar, and a foot pedal to generate the control parameters required to control the sound source and the filter bank. This system was displayed with great success at the New York World's Fair in 1939. According to Dudley, it took a few weeks of training to be able to operate a VODER and produce intelligible speech on demand.

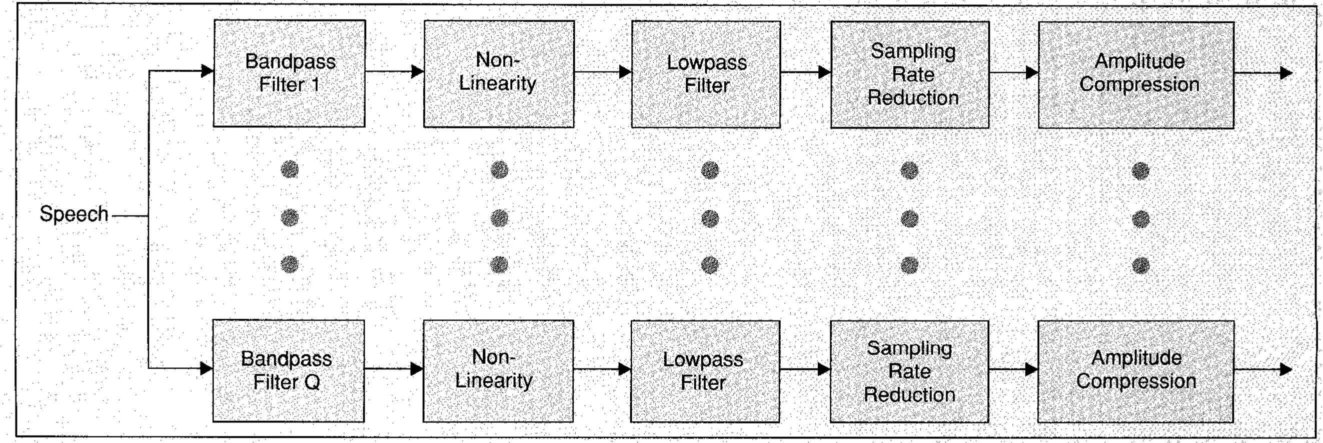
The VOCODER [4] derived its control parameters from a speech signal recorded using an attached microphone. From the speech signal the machine automatically determined the fundamental frequency (for voiced speech) as well as the gains for the bandpass filters. A value of zero for the fundamental frequency indicated that the hiss source was to be used. These control parameters, when used in the manner described above, produced a signal that was perceived to be similar to the input speech signal. It is worth noting that the *waveform* of the reconstructed signal generally was quite different from the waveform of the input signal. However, the time-variation of the distribution of speech energy with frequency was similar enough to fool the ear into judging the two signals to be similar in sound.

Dudley's demonstration that a speech signal could be represented in terms of a set of slowly varying parameters that could later be used as control parameters to resynthesize an approximately matching speech signal opened up the possibility of compressing the bandwidth of a speech signal. In modern digital telephony, this principle led to a series of methods for efficient digitization of speech for transmission (see the Speech Coding section).

From Tract Modeling to Spectral Estimation

Dudley used a bank of filters to control the sound spectrum in the VODER system. In order to produce the intended sound, the gain or attenuation of the filters at various frequencies had to be commensurate with the power of the input speech sound at those frequencies. Thus, the function of the filter bank was to model (nonparametrically) the vocal-tract response and, therefore, the need to measure proper attenuation values required sophisticated techniques for, in modern terms, spectral estimation [5]. The VOCODER, as proposed by Dudley, aimed at the very same purpose; namely, efficient estimation of the time-varying spectrum of the input speech signal.

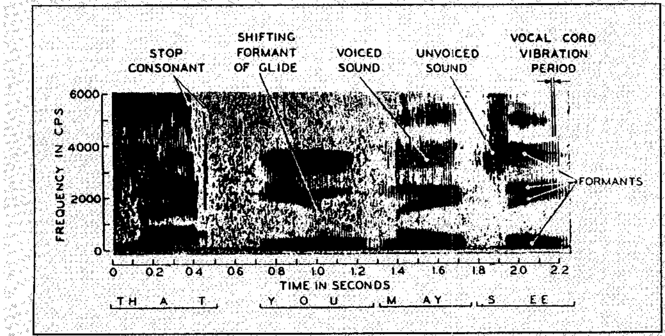
Spectral estimation using a filter bank (i.e., in essence the reverse of the VODER system) is depicted in Fig. 2. Each filter in the bank attempts to estimate the speech signal energy at and around the center frequency of the filter.



▲ *2. A block diagram for spectral estimation with filterbank.*

The nonlinearity that follows the filter measures the energy of the filtered signal. The result across the filter bank is an estimate of the spectral profile, or frequency response, that can be used to characterize the signal at a particular time.

The extraction of spectral control parameters from a speech signal has many other applications besides speech synthesis and bandwidth compression. Dudley himself realized that the pattern of variation of these parameters with time is characteristic of the utterance. This idea was explored by Dudley, and by many other researchers, for automatic recognition of speech by machine. The parameters could also be used to recognize the identity of a speaker from his or her voice. Finally the realization of the fundamental importance of these parameters led to the construction of the sound spectrograph [6] for displaying the time-varying spectra of speech (see Fig. 3). This in turn led to attempts at using the principles of the sound spectrograph (the sonogram) as a means for communication with the deaf, by teaching them how to recognize spoken words from displays of their time-varying spectra [7].



▲ *3. Broadband sound spectrogram of the utterance "That you may see synthesizer (after Dudley, Riesz, Watkins and Flanagan [2]).*

Linear Prediction

Representation of the vocal-tract frequency response, independent of the source parameters (e.g., voicing and fundamental frequency), captured researchers' interest in the 1960s. One approach to this problem was to analyze the speech signal using a transmission line analog of the wave-propagation equation. This method allows use of a time-varying source signal as excitation to the "linear" system of the vocal tract.

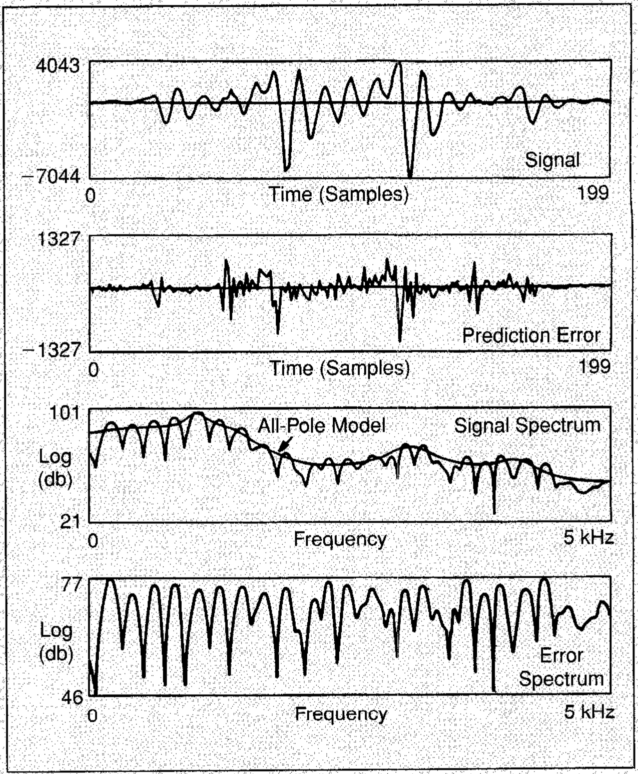
To make analysis of the vocal-tract response tractable, one often assumes that the vocal tract is an acoustic system consisting of a concatenation of uniform cylindrical sections of different areas with planar waves propagating through the system. Each section can be modeled with an equivalent circuit with wave reflections occurring at the junctions between sections. Such a model allows analysis of the system from its input-output characteristics [4, 8].

In the late 1960s, Atal [9] and Itakura [10] independently developed a spectral analysis method, now known as linear prediction. While the motivations were different, they made an identical assumption; namely, that the speech signal at time t could be approximately predicted by a linear combination of its past values. In a discrete time implementation of the method, this concept is expressed as

$$s_i \sim \hat{s}_i = \sum_{j=1}^{p} a_j s_{i-j}$$

where p is called the order of the analysis. The task is to find the coefficients $\{a_j\}$ that minimize some measure of the difference between s_i and $\hat{s}_i$ over a short-time analysis window. To retain the time-varying characteristics of the speech signal, the analysis procedure updates the coefficient estimation process progressively over time. This process is generally referred to as short-time spectral analysis.

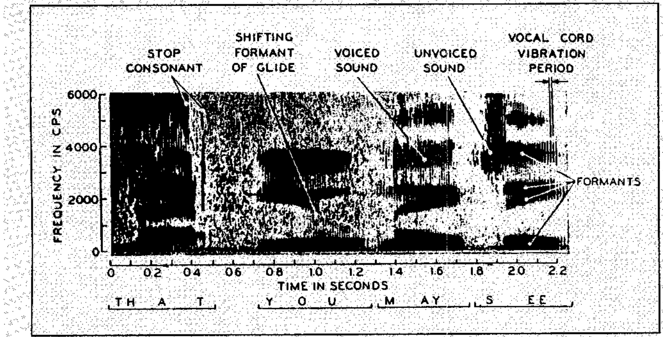
The linear prediction analysis method has several interesting interpretations. In the frequency domain, the computed coefficients $\{a_j\}$ define an all-pole spectrum $\sigma/A(e^{j\omega})$where $A(z) = 1 - \sum_{j=1}^{p} a_j z^{-1}$ with $z = e^{j\omega}$. Such a spectrum is essentially a short-term estimate of the spectral envelope of the speech signal, at a given time, as shown in Fig. 4. The "envelope" models the frequency response of the vocal tract while the fine structure in the Fourier spectrum is a manifestation of the source excitation or driving function. This spectral envelope estimate can be used for many purposes; e.g., as the spectral mag-



▲ *4. Typical signals and spectra for linear predictive coding (LPC) autocorrelation method for a segment of speech (after [89]).*

The nonlinearity that follows the filter measures the energy of the filtered signal. The result across the filter bank is an estimate of the spectral profile, or frequency response, that can be used to characterize the signal at a particular time.

The extraction of spectral control parameters from a speech signal has many other applications besides speech synthesis and bandwidth compression. Dudley himself realized that the pattern of variation of these parameters with time is characteristic of the utterance. This idea was explored by Dudley, and by many other researchers, for automatic recognition of speech by machine. The parameters could also be used to recognize the identity of a speaker from his or her voice. Finally the realization of the fundamental importance of these parameters led to the construction of the sound spectrograph [6] for displaying the time-varying spectra of speech (see Fig. 3). This in turn led to attempts at using the principles of the sound spectrograph (the sonogram) as a means for communication with the deaf, by teaching them how to recognize spoken words from displays of their time-varying spectra [7].



▲ *3. Broadband sound spectrogram of the utterance "That you may see synthesizer (after Dudley, Riesz, Watkins and Flanagan [2]).*

Linear Prediction

Representation of the vocal-tract frequency response, independent of the source parameters (e.g., voicing and fundamental frequency), captured researchers' interest in the 1960s. One approach to this problem was to analyze the speech signal using a transmission line analog of the wave-propagation equation. This method allows use of a time-varying source signal as excitation to the "linear" system of the vocal tract.

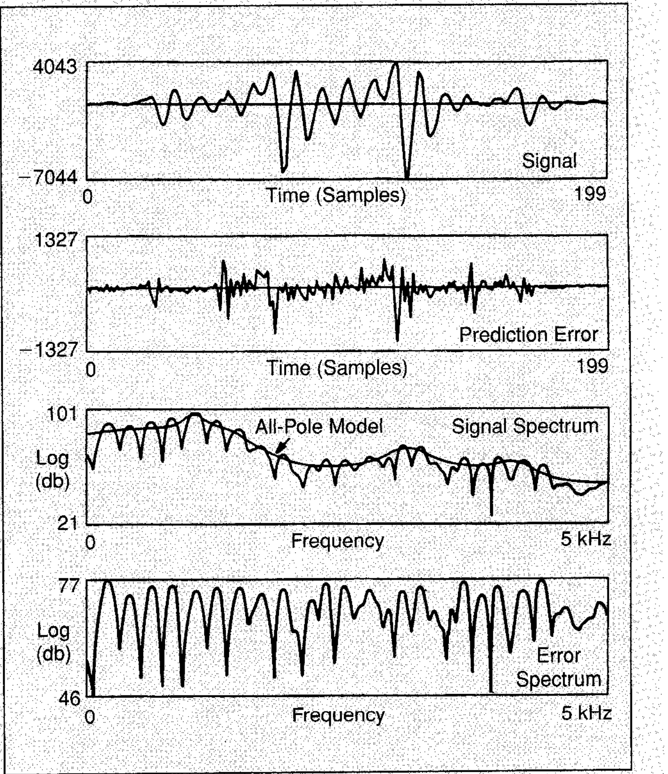
To make analysis of the vocal-tract response tractable, one often assumes that the vocal tract is an acoustic system consisting of a concatenation of uniform cylindrical sections of different areas with planar waves propagating through the system. Each section can be modeled with an equivalent circuit with wave reflections occurring at the junctions between sections. Such a model allows analysis of the system from its input-output characteristics [4, 8].

In the late 1960s, Atal [9] and Itakura [10] independently developed a spectral analysis method, now known as linear prediction. While the motivations were different, they made an identical assumption; namely, that the speech signal at time t could be approximately predicted by a linear combination of its past values. In a discrete time implementation of the method, this concept is expressed as

$$s_i \sim \hat{s}_i = \sum_{j=1}^{p} a_j s_{i-j}$$

where p is called the order of the analysis. The task is to find the coefficients $\{a_j\}$ that minimize some measure of the difference between s_i and $\hat{s}_i$ over a short-time analysis window. To retain the time-varying characteristics of the speech signal, the analysis procedure updates the coefficient estimation process progressively over time. This process is generally referred to as short-time spectral analysis.

The linear prediction analysis method has several interesting interpretations. In the frequency domain, the computed coefficients $\{a_j\}$ define an all-pole spectrum $\sigma/A(e^{j\omega})$where $A(z) = 1 - \sum_{j=1}^{p} a_j z^{-1}$ with $z = e^{j\omega}$. Such a spectrum is essentially a short-term estimate of the spectral envelope of the speech signal, at a given time, as shown in Fig. 4. The "envelope" models the frequency response of the vocal tract while the fine structure in the Fourier spectrum is a manifestation of the source excitation or driving function. This spectral envelope estimate can be used for many purposes; e.g., as the spectral mag-



▲ *4. Typical signals and spectra for linear predictive coding (LPC) autocorrelation method for a segment of speech (after [89]).*

analog networks were serial or parallel combinations of second-order resonators. A series of impulse-like waveforms, or white noise, was applied to the resonators in order to generate vowels or fricative sounds.

In the 1960s, the discrete domain realizations of formant synthesizers were proposed [8, 34]. The resonators for the formant synthesizer were arranged in either a cascade or parallel manner [8,35,36]. Flanagan concluded that the serial form was a better model for non-nasal voiced sounds, while the parallel structure was superior for nasal and unvoiced sounds. The reason was that the vocal tract is considered as an all-pole filter for non-nasal voiced sounds and as a pole-zero system for other phonations. Thus, it is quite simple to use the cascade structure to simulate an all-pole system and the parallel form to implement a pole-zero system. Klatt's system combined the cascade and the parallel structures. Anti-resonances were added to the cascade branch to enhance the ability of the cascade configuration to model nasal and unvoiced sounds. When the synthesis variables are properly specified and the correct configuration is used, this synthesizer is capable of synthesizing high-quality, intelligible speech [37].

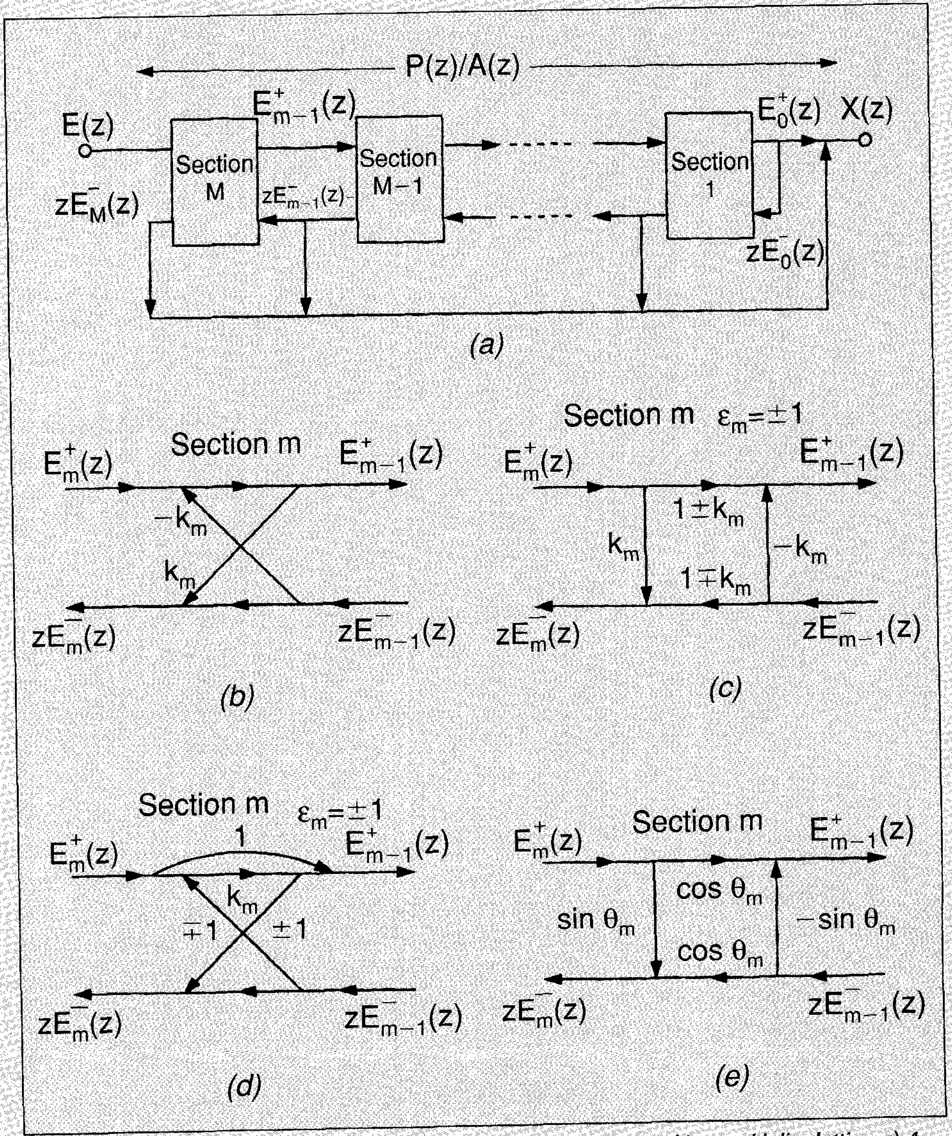
Linear Prediction (LP) synthesis

The linear predictive synthesizer is a mathematical all-pole realization of the linear source-tract model [9]. The linear prediction all-pole filter is an IIR (infinite impulse response) filter, and a wide range of structures were proposed for digital implementation of linear prediction synthesizers [38, 39].

Aside from the usual digital filter implementations (direct form, parallel form, cascade form, etc.), structures developed for linear prediction synthesis include: 1) a 2-multiplier lattice; 2) a 4-multiplier ladder (having the form of the Kelly–Lochbaum model [21]); 3) a 1-multiplier form; and 4) a 4-multiplier normalized form [40]. These implementations are shown in Fig. 5. These structures were developed for two major reasons: (a) they allow the synthesis filter to be implemented directly from the reflection coefficients, and (b) in an actual computer implementation, they allow one to trade off accuracy, the number of multiplications and additions, and complexity [40]. These are important considerations in the realization of synthesis technology.

Related Topics

In the early and mid 1980s, Hanson et al. [41] as well as McAulay and Quatieri [42] developed a sinusoidal model for speech analysis/synthesis. This method has found use for speech transformations, such as time-scale and pitch-scale modifications. Molines and Charpentier [43] suggested the pitch-synchronous overlap-add (PSOLA) approach for text-to-speech applications. This approach can modify the prosody of the speech and is able to concatenate speech waveforms. The speech is modified in either the time domain or the frequency domain. Other applications of speech synthesis include reading e-mail, fax, and webpages, and as a proofing tool for previewing text in word processors.



▲ *5. Various forms of digital filter implementation: a) general form; b) 2-multiplier lattice; c) 4-multiplier ladder d) 1-multiplier form; e) 4-multiplier normalized form.*

Speech Coding

Homer Dudley's pioneering work [3] was motivated by the need to increase the communication capacity (number of channels) in a telephone network (which was analog then). The term "bandwidth compression" was generally used to refer to such a task. Today, most if not all of the telephone network is digital and, hence, speech bandwidth compression translates into speech coding, which aims at representing the speech signal in binary digits (bits) with highest efficiency (i.e., highest quality of the reconstructed signal with least number of bits).

Digital encoding of speech begins with an analog-to-digital conversion device that samples the analog speech waveform at an appropriate rate (usually 8,000 samples per second for telephone bandwidth speech) and then represents the amplitude of each sample digitally. In communication systems, this is the so-called pulse-coded modulation (PCM). Typically, each waveform sample is represented by 12-16 bits, resulting in a rate of 96-128 thousand bits per second (kbps or kb/s). Research in speech coding attempts to find methods to increase the efficiency in transmission and storage while maintaining the speech quality.

Aside from efficient transmission, speech coding is also essential for achieving secure communications. This is the main reason that speech compression and coding research benefited from strong government support in the past five decades. The "A History of Secure Voice Coding" sidebar presented with this article provides a brief, chronological perspective of this work.

In general, speech-coder attributes can be described in terms of four classes: *bit rate, complexity, delay, and quality*. The *bit rate* is the communication channel bandwidth at which the coder operates. Digital network telephony generally operates at 64 kb/s, cellular systems operate from 6.7 to 13 kb/s, and secure telephony at 2.4 and 4.8 kb/s. Systems can also be designed to take advantage of the natural silences that take place during speech. CDMA digital cellular telephony employs variable-rate speech coders that operate at maximum rate during a talk-spurt and minimal rate during silence.

Complexity refers to the computational complexity of the speech coder. For most applications, speech coders are implemented on either special-purpose devices (such as DSP chips) or on general-purpose computers (such as a PC for Internet telephony). In either case, the important quantities are the number of (million) instructions per second that are needed to operate in real-time and the amount of memory used. The greater the memory usage and the greater the number of instructions per second, the more expensive and power consuming the implementation platform. This has important consequences for most applications.

Delay refers to the communications delay caused by the coder. One component of the delay is due to the algorithm and the other to the computation time. Individual sample coders have the lowest delay, while coders that work on a block or frame of samples have greater delay. Too much delay can have serious repercussions on a conversation. Excessive delay creates critical challenges on the network echo canceler and also forces speakers into an inconvenient "push-to-talk" mode, making conversation ineffective. The practical limit of round-trip delay for telephony is about 300 ms. With the advent of packet telephony, other sources of delay may be present, affecting the design of the speech coder.

Quality refers to a large number of attributes. As bit rates are lowered, speech coders become more speech specific and give less-faithful renditions of other sounds. While music can be transmitted through 64 kb/s PCM, it may be unrecognizable over some 2.4 kb/s coders. Background noises such as babble, traffic noise, or noise inside a car, office, shopping mall, etc., can all affect the perceived quality of a speech coder. For many applications, speech coders are tandemed. For example, accessing a voicemail system from a cellular phone may involve two different encodings. Quality and even intelligibility may suffer.

In selecting a speech coder for a given application, the designer can make tradeoffs among these four classes of attributes.

Today, speech coding finds a diverse range of applications such as cellular telephony, voice mail, multimedia messaging, digital answering machines, packet telephony, audio-visual teleconferencing, and of course many other applications in the Internet arena.

From Quantization to Model-Based Coding

Digital representation of a signal requires quantization of the amplitude; i.e., an analog sample of infinite precision needs to be converted to a discrete number that can be represented by a fixed number of bits. This is the first step in speech coding. Early research focused on the design of a quantization table (the set of values used to represent speech) that minimizes the average quantization noise (discrepancy between the original value and the represented one) [44-47]. Signal companding (compression and expansion) [48] such as μ-law or A-law is often used to transform the signal statistics (on a sample by sample basis) for improved coding efficiency [44]. In digital telephony, μ-law and A-law PCM [44, 48] are the schemes that were adopted for transmitting speech at 64 kbps (or 56 kbps).

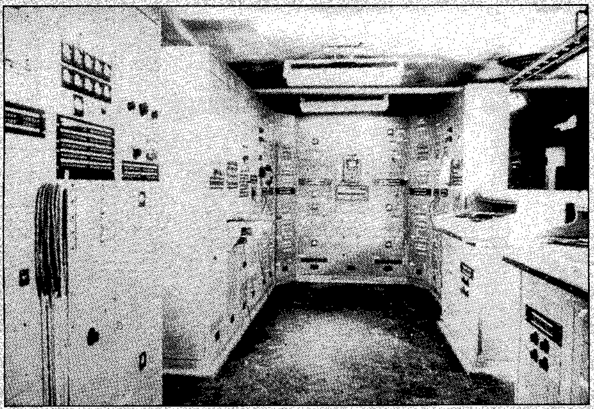
Minimization of quantization noise requires critical knowledge of the signal statistics. Since speech characteristics vary with time, improvements (further reduction of quantization noise) can be achieved by adaptive quantizers [49, 50], which adjust the quantization table according to the time-varying signal properties. Adaptation can be implemented in either a forward or backward manner (or in more sophisticated systems, a combination of both) [44].

A History of Secure Voice Coding

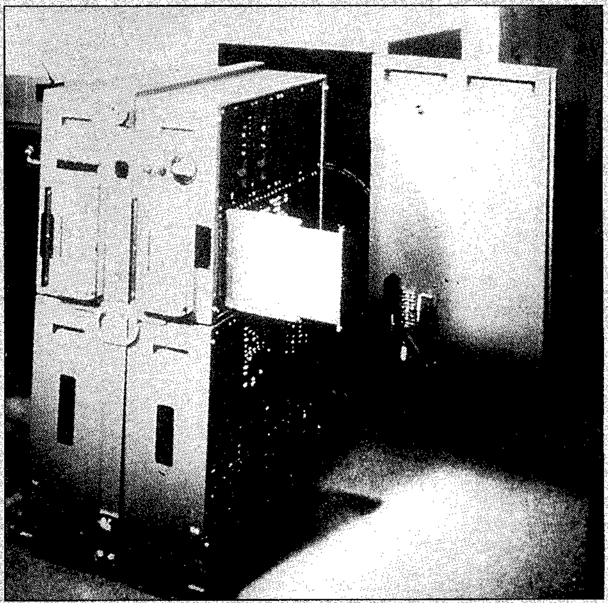
Joseph P. Campbell Jr. and Richard A. Dean

The NSA (National Security Agency of the U.S. Government) inherited the responsibilities, traditions, and expertise of voice coding from the Army Security Agency for enciphered speech applications in 1952. The engineers and technicians who had participated in the now famous SIGSALY vocoder used by Roosevelt and Churchill for planning the "D-Day" invasion were still developing their craft at NSA. SIGSALY, shown in Fig. A1, was a vocoder-based system not unlike the "Talking Machine" first introduced by Homer Dudley of Bell Labs at the 1939 World's Fair. Developed with Bell Labs, it consisted of a bank of 10 bandpass filters spaced approximately at the bands of equal articulation for speech, from baseband up to 3000 Hz. Each filter could be excited by one of six logarithmically spaced amplitudes as developed by Harry Nyquist in the first application of PCM. A "Buzz"/"Hiss" generator was used as an exciter for the vocoder corresponding to the voiced/unvoiced attribute of each 20 msec speech segment. Balance of the "Buzz"/"Hiss" generator, or voicing, represented a major factor in the quality of the speech. To test and tune this delicate balance, Mitchell Brown developed the "Aaahhh"-"Sshhhhh" test to check for the proper balance of these sounds.

Tom Tremain of NSA expanded upon vocoder testing and he instituted a defacto standard in diagnostic testing. His unique ability to interpret diagnostic rhyme tests (DRTs) and diagnostic acceptability measures (DAMs) to pinpoint weaknesses in speech-coding algorithms and systems led researchers from around the world to seek his expert advice to improve their systems.



▲ *A1. A view of SIGSALY through the rear doors of a tractor trailer. The first real-time secure voice system, circa 1944.*



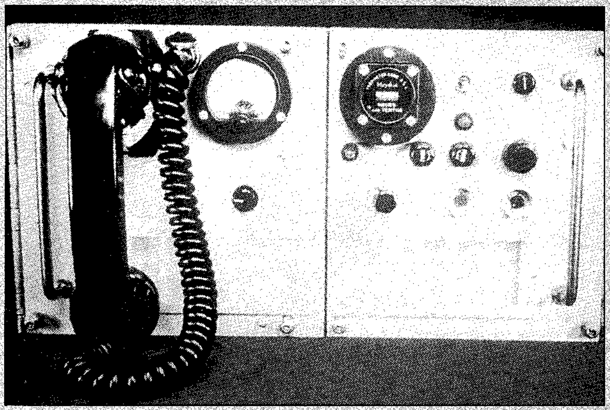
▲ *A2. The KY-9 transistorized secure voice system, circa 1953.*

From the time of SIGSALY until the beginning of 1960s, several generations of voice coders had been developed in conjunction with Bell Labs. The KO-6 voice coder, developed in 1949 and deployed in limited quantity, was a close approximation to the 1200 bps SIGSALY voice coder. This was followed in 1953 by the 1650 bps KY-9 shown in Fig. A2. It used a 12-channel vocoder and hand-made transistors and was one of the earliest applications of solid-state technology. This resulted in a remarkable weight reduction of SIGSALY's vacuum tube technology (from 55 tons to a mere 565 pounds!). In 1961, Tremain's first project

continued on next page

The speech signal (due to its generation in an articulatory process) typically has a low-pass characteristic with roughly a 6 dB/octave roll-off. This property is the basis of a differential quantization scheme that encodes the difference between successive samples rather than the original sample value. The differentiator essentially equalizes the long-term speech spectrum (makes it flat across frequency) and reduces the signal variance for easier quantization. The method is generally referred to as differential PCM (DPCM) [49, 51] coding. When the coefficient of the differentiator and the quantization table are made adaptive to the local signal characteristics, it is called adaptive DPCM (ADPCM) [52, 53].

A differential coding scheme can be further elaborated; rather than coding the difference between successive samples, it can code the output of a higher-order filter involving a fixed number of past sample values. The scheme then becomes that of adaptive predictive coding (APC) [54], which shares a similar interpretation to linear predictive coding (LPC) [9] in terms of vocal-tract response modeling. That is, the predictor filter tracks the time-varying characteristics of the vocal tract. The effect of prediction in coding is reduction of signal variance (the prediction error signal or residual has a smaller variance than that of the original signal) and whitening of the signal spectrum (the



▲ *A3. The HY-2 channel vocoder, circa 1961.*

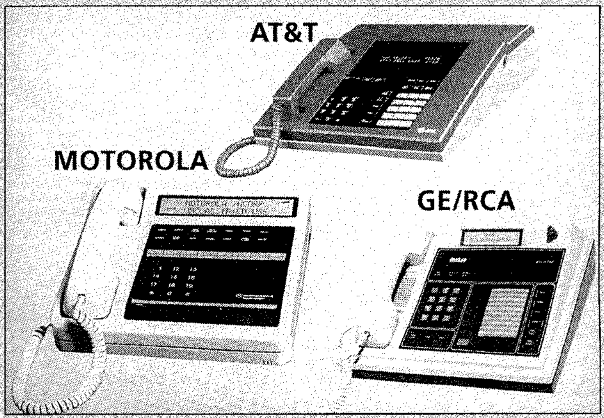
at NSA, led by Brown, was the development of the HY-2 vocoder, the last generation of U.S. channel vocoder technology. The HY-2, shown in Fig. A3, was a 16-channel 2400-bps system using "Flyball" color-coded modular logic to reduce the weight down to 100 pounds.

Even the best of U.S. vocoder technology was limited by the analog technology that was the basis of its implementation. As the analog filters and amplifiers vary with age and temperature, so does the sensitive tracking required between the speech analyzer and speech synthesizer. Performance in the field never approached laboratory performance, and users, starting with Churchill and Roosevelt, were reluctant to use systems that had a synthetic "Donald Duck" quality. President Johnson refused to use the HY-2 because of its poor quality and, as a result, deployment was limited.

In the late 1960s, digital signal processing for voice was considered a productive new direction. Working with Bell Labs, NSA started the linear predictive coding (LPC) generation of voice coders. In 1974, the first real-time computer simulation of LPC-10 on the CSP-30 computer was demonstrated, which was a milestone in signal-processing history.

NSA's real-time digital voice coder led to a whole new family of NSA vocoder products, called secure telephone units (STUs), built around the first generation of AMD2901-based high speed bit-slice signal processors that forever changed the way voice coding was accomplished. Today's STU-III, shown in Fig. A4, is the third-generation desktop telephone that uses an enhanced LPC-10 and supports secure voice users throughout the government. Tremain's team, in cooperation with Bell Labs, developed the Federal Standard 1016 Code Excited Linear Prediction (CELP) speech coder in the late 1980s. Through Tremain's efforts, the CELP coder was developed and deployed in later models of the STU-III and saw successful variations adopted into cellular phones. Before his death in 1995, Tremain was also pivotal in brining about the new 2400 bps vocoder based on MELP.

Voice coders, long associated only with exotic encryption schemes, are finding numerous applications today for wireless communications, voice mail, network communications, and synthetic voice applications. Today's voice coders used in satellites, answering machines, talking toys, audio over the world wide web, Internet phones, and hand-held digital cellular and microcellular telephones are just a few of the direct descendants of the secure voice-coding work.

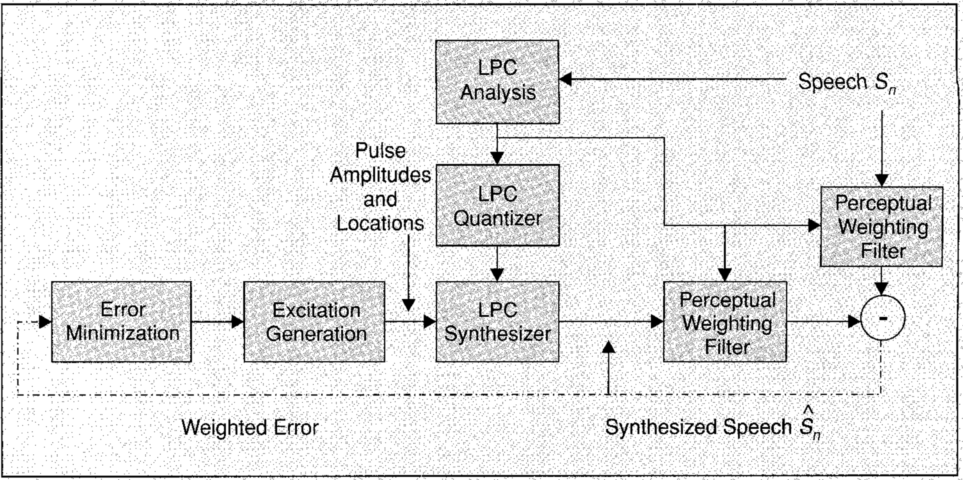


▲ *A4. The STU-III secure voice terminal family, circa 1986.*

error signal is essentially uncorrelated since most of the signal redundancy is represented by the predictor coefficients).

In the 1970s, researchers started to explore the possibility of incorporating our perceptual knowledge of auditory masking in coding schemes, in addition to attempting to invent new coding structures. Atal and Schroeder [55] proposed the concept of error signal shaping with the implication that the coding error can be made imperceptible (masked by the coded signal) if its spectrum is properly shaped and stays below the audible threshold in the presence of the co-existing signal. This concept led to the use of perceptual weighting in the error criterion used by most of the analysis-by-synthesis coding structures [56]. The same concept has also been used in bit-allocation schemes [57].

Figure 6 is a block diagram of a generic analysis-by-synthesis coding structure. The speech is first analyzed to obtain the LPC synthesis filter for a frame of speech. A perceptual weighting filter is derived from the LPC filter. The speech is passed through the perceptual weighting filter to form the target signal. The possible excitation sequences are passed through the combination of the LPC filter and perceptual weighting filter. The excitation signal that minimizes the mean square error (MSE) between the weighted output signal and the target signal is selected. The pitch properties of the speech signal can be exploited prior to selecting the excitation.


LPC Analysis
Speech S_n
Pulse Amplitudes and Locations
LPC Quantizer
Perceptual Weighting Filter
Error Minimization
Excitation Generation
LPC Synthesizer
Perceptual Weighting Filter
-
Weighted Error
Synthesized Speech $\hat{S}_n$

▲ 6. *A block diagram of a general analysis-by-synthesis coding structure.*

Analysis-by-synthesis coders are essentially waveform-approximating coders because they produce an output waveform that follows closely the original waveform. (The minimization of the MSE in the perceptual space via perceptual weighting causes a slight modification to the waveform-approximation principle.) This avoids the old vocoder problem of classifying a speech segment as voiced or unvoiced. Such decisions can never be made flawlessly and many speech segments have both voiced and unvoiced properties.

Today's vocoders also have found ways to avoid making the voiced/unvoiced decision. The multiband excitation (MBE) [58] and sinusoidal transform coders (STC) [42], also known as harmonic coders, divide the spectrum into a set of harmonic bands. Individual bands can be declared voiced or unvoiced. This allows the coder to produce a mixed signal: partially voiced and partially unvoiced. Mixed-excitation LPC (MELP) [59] and waveform interpolation (WI) [60] produce excitation signals that are a combination of periodic and noise-like components. These modern vocoders produce excellent-quality speech compared to their predecessors, the channel vocoder [61] and the LPC vocoder [62]. However, they are still less robust than higher-bit-rate waveform coders. They are more affected by background noise and cannot code music well.

Vector Quantization

Advances in coding theory suggest that optimal coding efficiency can be attained asymptotically as the number of signal samples encoded simultaneously is increased [63]. This motivated speech-coding researchers in the late 1970s and 1980s to explore the use of the methods of vector quantization (as opposed to scalar, or single sample) schemes.

Vector quantization aims at encoding an entire vector of samples or coefficients simultaneously. The technique was applied to spectral-parameter [64, 65] as well as to waveform quantization [66]. Today, vector quantization is used in most speech coders.

Research in vector quantization focused on methods for generating the codebook [67], the type of distortion measures [64], and efficient structures to achieve high-rate, low-distortion VQ [68]. Vector quantization was also essential in achieving extremely low-bit-rate (less than 1000 bps) vocoders [65].

Speech-Coding Standards

For speech coding to be useful in telecommunication applications, it has to be standardized (i.e., it must conform to the same algorithm and bit format) to ensure universal interoperability. Speech-coding standards are established by various standards organizations: for example, the International Telecommunications Union (ITU), the Telecommunications Industry Association (TIA), the Research and Development Center for Radio Systems (RCR) in Japan, the International Maritime Satellite Corporation (Inmarsat), the European Telecommunications Standards Institute (ETSI), and other government agencies.

The ITU (formerly CCITT) defined the "first" speech-coding algorithm for digital telephony in 1972. It is the 64 kb/s companded PCM coder. In North America and Japan, μ-law PCM is used. In the rest of the world, A-law PCM is used. These coders use 8 bits to represent each sample of the speech signal with a sampling rate of 8 kHz (i.e., maximum signal frequency of 4 kHz). The standard is referred to as G.711 [69].

In 1984, Recommendation G.721 [70], which is based on ADPCM coding operating at 32 kb/s, was standardized for digital circuit multiplication equipment. Associated with G.721 were 1) G.723 [69], which extends G.721 to two additional bit rates, 24 and 40 kb/s; 2) G.726 [69], which unifies and replaces G.721 and G.723 and extends it to 16 kb/s; 3) G.727 [69], which has an even number of levels for all associated coders.

The low-delay, code-excited-linear-prediction (LD-CELP) coder was standardized in 1992 and 1994 for 16 kb/s applications. It is designated as Recommendation G.728 [71]. Furthermore, G.729 (8 kb/s) and G.723.1 (5.3 and 6.3 kb/s) were subsequently standardized in 1995. Both coders are based on the analysis-by-synthesis structure. For wideband (7 kHz bandwidth) speech, Recommendation G.722 [72] was established in 1988 for bit rates of 48, 56 and 64 kb/s.

For digital cellular applications, the European Groupe Special Mobile (GSM) of CEPT defined a 13kb/s coder in

1987 based on the regular-pulse-excitation with long-term-predictor (RPE–LTP) coding algorithm [73]. Another coder defined by ETSI in 1994 was the 5.6 kb/s vector-sum-excited-linear-prediction (VSELP) coder [74], known as GSM Half-Rate. In North America, VSELP was also adopted in 1989 as the TIA IS54 [54] coder at 8 kb/s (7.95 kb/s) for digital cellular telephony. In 1993, IS96 [75], the CELP-based coder, was recommended for CDMA cellular systems operating at bit rates 8.0, 4, 2, and 0.8 kb/s. Most recently, IS-641 was recommended as an improved coder at 8 kb/s for TDMA cellular systems and IS-127 (or EVRC, enhanced variable bit-rate coder) for CDMA applications.

Finally, the U.S. Department of Defense (DoD) announced FS1015 [76] based on linear prediction as the standard coder at 2.4 kb/s for secure voice applications in 1984. In 1991, the DoD further adopted a CELP based coder at 4.8 kb/s as the FS1016 standard [77]. A new 2.4 kb/s coder based on MELP was announced in 1996 at ICASSP in a session dedicated to Tom Tremain [59].

New Challenges

Most of the low-bit speech coders designed in the past implicitly assume that the signal is generated by a speaker without much interference. These coders often demonstrate degradation in quality when used in an environment in which there is a competing speech or background noise. A recent research challenge is to make coders perform robustly under a wide range of conditions, including noisy automobile environments.

Another challenge is the coder's resistance to transmission errors, which are particularly critical in cellular and packet communication applications. Methods that combine source and channel coding schemes or conceal errors are important in enhancing the usefulness of the coding system.

As packet networking is becoming more and more prevalent, a new breed of speech coders is emerging. These coders need to take into account and negotiate for the available network resources (unlike the existing digital telephony hierarchy in which a constant bit rate per channel is guaranteed) in order to determine the right coder to use. They also have to be able to deal with packet losses (severe at times). For this reason, the idea of embedded and scaleable (in terms of bit rates) coders is being investigated, with much interest [78].

Speech Enhancement

The idea that vocoder principles could be used to improve the quality of a speech signal corrupted by additive noise was first introduced by M.R. Schroeder in 1960 [79]. The basic idea was to generate a signal with a fine structure as close as possible to that of the original speech signal, but with an envelope that attenuates the signal between formant peaks. This idea, with several modifications, was first simulated by Sievers and Sondhi [80] in 1964. Although the idea was shown to be feasible, the quality attained was not very good.

Since those early days, variants of this idea have been proposed and implemented by several authors, notably Weiss, Aschkenasy, and Parsons [81]; Boll [82], McAulay and Malpass [83]; Ephraim and Malah [84]; and Lim and Oppenheim [85]. The common features of all these implementations are to split the noisy speech signal into frequency regions by passing it through a filter bank and attenuating the output of each channel by a factor depending on the estimated signal-to-noise ratio in that channel. The main differences between these various proposals are the methods used to estimate the level of noise and of speech in various frequency bands.

A method proposed by Ephraim, Malah, and Juang [86] might formally be classified as belonging to this category. However, it differs from the rest in that it bases its selective attenuation of the various frequencies on hidden Markov models (HMMs) of the noise and the speech.

Enhancement of speech signals in noise has been quite useful in telephony applications. Some recent implementations of Etter [87] and Diethorn [88] are some of the best examples of this application.

Speech Recognition

Speech recognition by machine in a limited and strict sense can be considered as a problem of converting a speech waveform into words. It requires analysis of the speech signal, conversion of the signal into elementary units of speech such as phonemes or words, and interpretation of the converted sequence in order to allow correction of the misrecognized words/units or for other linguistic processing such as parsing and speech understanding.

A Brief History of the Research (after [89])

Research in automatic speech recognition by machine has been done for almost four decades. The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [90]. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words [91]. The system again relied on spectral measurements (as provided by an analog filter bank) primarily during vowel regions. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [92]. They used a spectrum analyzer and a pattern matcher to make the recognition decision. A novel aspect of this research was the use of statistical information about allowable sequences of phonemes in English (a rudimentary form of language syntax) to improve overall phoneme accuracy for words consisting of two or more phonemes. Another effort of note in this period was the vowel recognizer of Forgie and Forgie, constructed at MIT Lincoln Laboratories in 1959, in which 10 vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker-independent manner [93]. Again a filter-bank analyzer was used to provide spectral information, and a time-varying estimate of the vocal-tract resonances was made to decide which vowel was spoken.

In the 1960s several fundamental ideas in speech recognition surfaced and were published. However, the decade started with several Japanese laboratories entering the recognition arena and building special-purpose hardware as part of their systems. One early Japanese system, described by Suzuki and Nakata of the Radio Research Lab in Tokyo [94], was a hardware vowel recognizer. An elaborate filter-bank spectrum analyzer was used along with logic that connected the outputs of each channel of the spectrum analyzer (in a weighted manner) to a vowel-decision circuit, and a majority-decision logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita of Kyoto University in 1962, who built a hardware phoneme recognizer [95]. A hardware speech segmenter was used along with a zero-crossing analysis of different regions of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963 [96]. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program.

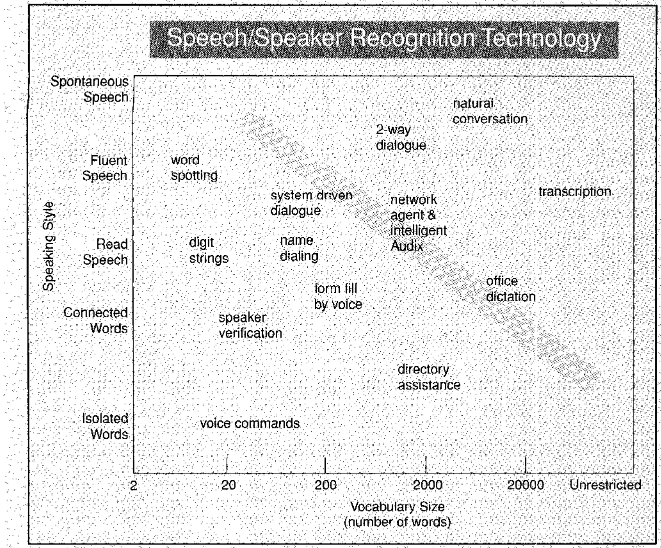
In the 1960s three key research projects were initiated that have had major implications on the research and development of speech recognition for the past 20 years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems associated with nonuniformity of time scales in speech events. Martin developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduced the variability of the recognition scores [97]. Martin ultimately developed the method and founded one of the first companies, Threshold Technology, which built, marketed, and sold speech-recognition products. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances [98]. Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected-word recognition, were embodied in Vintsyuk's work, it was largely unknown in the West and did not come to light until the early 1980s; this was long after the more formal methods were proposed and implemented by others.

A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes [99]. Reddy's research eventually spawned a long and highly successful speech-recognition research program at Carnegie Mellon University (CMU) (to which Reddy moved in the late 1960s). One of the first demonstrations of spoken-language understanding at CMU was in 1973. The Hearsay I System, developed at CMU, was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. In the Voice Chess task domain used by Hearsay I, the number of alternative sentences that could be spoken at any given point was limited to the synonyms of the possible moves. There are not yet many systems that effectively demonstrate the role of semantics in reducing the complexity of search. However, the principle that syntactic, semantic, and contextual knowledge sources can be used to reduce the number of possible alternatives to be considered in decoding appears to be central to the design of spoken-language-understanding systems.

In the 1970s speech-recognition research achieved a number of significant milestones. First, the area of isolated-word or discrete-utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in the Soviet Union [100], Sakoe and Chiba in Japan [101], and Itakura in the United States [102]. The Russian studies helped advance the use of pattern-recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura's research showed how the ideas of LPC, which had already been successfully used in low-bit-rate speech coding, could be extended to speech-recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.

Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large-vocabulary automatic speech dictation at IBM in which researchers studied three distinct tasks over a period of almost two decades (namely, the New Raleigh language [103] for simple database queries, the laser patent text language [104] for transcribing laser patents, and the office correspondence task) with a system called Tangora [104], for dictation of simple memos.

Finally, at AT&T Bell Labs (now Bell Labs, Lucent Technologies, and AT&T Labs–Research), researchers began a series of experiments aimed at making speech-recognition systems that were truly speaker-independent [106] for telecommunication applications. The intended application was telecommunication services, where humans and machines conduct dialogues in order to accomplish a task such as routing a call, or making a reservation on cars or flights. To achieve this goal, a wide range of sophisticated algorithms were developed to deal with all variations of different words and different expressions across a wide user population. This research has been re-



▲ *7. Dimensions of automatic-speech-recognition applications and the current capabilities (shaded line).*

fined over a decade so that the techniques for creating speaker-independent speech models are now well understood and widely used.

Just as isolated word recognition was a key focus of research in the 1970s, the problem of connected-word recognition was a focus of research in the 1980s. Here the goal was to create a robust system capable of recognizing a fluently spoken string of words (e.g., digits) based on matching a concatenated pattern of individual words. A wide variety of connected-word-recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) [107], the one-pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in England [108], the level-building approach of Myers and Rabiner at Bell Labs [109], and the frame-synchronous level-building approach of Lee and Rabiner at Bell Labs [110]. Each of these "optimal" matching procedures had its own implementational advantages, which were exploited for a wide range of tasks.

Speech research in the 1980s was characterized by a shift in technology from template-based approaches to statistical modeling methods—especially the HMM approach [111, 112] (discussed later).

The success of hidden Markov modeling gave rise to a major impetus in the 1980s to large-vocabulary, continuous-speech-recognition systems by the Defense Advanced Research Projects Agency (DARPA) community. (For ARPA efforts in speech understanding in the 1970s, see [113].) Major research contributions resulted from efforts at CMU (notably the well-known SPHINX system) [114], BBN with the BYBLOS system [115], Lincoln Labs [116], SRI [117], MIT [118], and AT&T Bell Labs [119]. The DARPA program has continued into the 1990s, with emphasis shifting from air-travel information retrieval to a range of different speech-understanding applications areas, in conjunction with a new focus on transcription of broadcast news. At the same time, speech-recognition technology has been increasingly used within traditional telecom networks to automate as well as enhance operator services [120]. Figure 7 shows a plot of various applications of speech-recognition technologies along the dimensions of vocabulary size and speaking style. The level of difficulty increases roughly along the diagonal line away from the lower-left corner, and the shaded bar represents a threshold of applications that can be supported by the current technology. Many challenges are still ahead of us.

From Speech Analysis to Statistical Modeling

Until the 1970s and 1980s, automatic speech recognition was mostly considered to be a speech-analysis problem. The fundamental belief was that if a proper analysis method were available that could reliably produce the identity of a speech sound, recognition of speech would be readily attainable. Such a deterministic view of the speech-recognition problem was advocated by researchers in acoustic-phonetics by citing such examples as "A stitch in dime saves nine" (in contrast to "A stitch in time saves nine"), which they believe can only be recognized correctly by deriving acoustic-phonetic features. This view may be appropriate in a microscopic sense but does not address the macroscopic question of how a recognizer should be designed such that, on average (in dealing with all the input sounds), it achieves the least errors or error rate. Similarly, template-matching in most practical systems without a proper statistical foundation does not provide a rigorous answer to this question, which is best addressed by Bayes' decision theory. (Template-matching with *asymptotically* dense reference patterns certainly would fall into the category of nonparametric statistical-pattern-recognition approaches whose optimality can be analyzed in reference to the Bayes decision theory formulation.)

Bayes Decision Theory

Bayes decision theory deals with random observations from an information source consisting of M classes of events where the goal is to identify which class of event the observation belongs to. Let the joint probability of X (the observation) and C_i (the class identity), $P(X, C_i)$, be known to the designer of the classifier. In other words, the designer has full knowledge of the random nature of the source. To measure the performance of the classifier, for every class pair (j,i), a cost or loss function, e_{ji}, is defined to signify the cost of classifying (or recognizing) an observation from class i as belonging to a class j event. The loss function is generally nonnegative with $e_{ii} = 0$ representing correct classification.

Given an arbitrary observation, X, a conditional loss for classifying X as belonging to a class, I, event can be defined as

$$R(C_i|X) = \sum_{j=1}^{M} e_{ji} P(C_j|X) \tag{1}$$

where $P(C_j \mid X)$ is the *a posteriori* probability. This leads to a reasonable performance measure for the classifier, i.e. the expected loss, defined as

$$\mathcal{L} = \int R(C(X)|X)p(X)dX \tag{2}$$

where $C(X)$ represents the classifier's decision, assuming one of the M "values," $C_1, C_2 \ldots C_M$.

For speech recognition, the loss function, e_{ij}, is usually chosen to be the zero-one loss function defined by $e_{ij}=0$ for $i = j$ and $=1$ for $i \neq j$, $i, j = 1,2 \ldots M$, which assigns no loss to correct classification and a unit loss to any error, regardless of the class. With this type of loss function, the expected loss, $\mathcal{L}$, is, thus, the error probability of classification or recognition. The conditional loss becomes

$$\begin{aligned} R(C_i|X) &= \sum_{i \neq j} P(C_j|X) \\ &= 1 - P(C_i|X) \end{aligned} \tag{3}$$

The optimal classifier that achieves minimum $\mathcal{L}$ is thus the one that implements the following:

$$C(X) = C_i \quad \text{if} \quad P(C_i|X) = \max_j P(C_j|X). \tag{4}$$

In other words, for minimum-error-rate classification, the classifier employs the decision rule of Eq. (4), which is called the *maximum a posteriori* (MAP) decision. The minimum error achieved by the MAP decision is called *Bayes risk*. (It's worth being somewhat mathematical here since formulating the recognizer's performance in terms of minimum expected loss is the basis of the paradigm shift from deterministic pattern matching to statistical-pattern recognition.)

The required knowledge for an optimal classification decision is, thus, the *a posteriori* probabilities for the implementation of the MAP rule. These probabilities, however, are *not* known in advance and generally have to be estimated from a training data set with known class labels. Bayes decision theory thus effectively transforms the classifier design problem into a distribution estimation problem. This is the basis of the statistical approach to pattern recognition.

The *a posteriori* probability $P(C_i \mid X)$ can be rewritten as

$$P(C_i|X) = P(X|C_i)P(C_i) / P(X).$$

Since $P(X)$ is not a function of the class index and, thus, has no effect in the MAP decision, the needed probabilistic knowledge can be represented by the class prior, $P(C_i)$, and the conditional probability $P(X \mid C_i)$.

Probability Distributions for Speech

The statistical method, as discussed above, requires that a proper, usually parametric, distribution form for the observations be chosen in order to implement the MAP decision. A key issue is what is the right distribution form for speech utterances? This question involves two essential aspects: i) finding the speech dimensions that carry the most pertinent linguistic information, and ii) deciding how to statistically characterize the information along the chosen dimensions.

Based on empirical observations, the HMM was proposed [15, 121, 122] as a simple means to characterize speech signals. For detailed discussions of the HMM, references [112] and [122] provide good insights.

Developments of HMM

The statistical method of hidden Markov modeling for speech recognition encompasses several interesting problems, particularly the estimation problem [111, 123, 124, 125]. Given an observation sequence (or a set of sequences), X, the estimation problem involves finding the "right" model parameter values that specify a source model (probability distribution) most likely to produce the given sequence of observations. In solving the estimation problem, we usually use the method of maximum likelihood (ML); that is, we choose λ such that $P(X \mid \lambda)$ is maximized for the given "training" sequence, X.

Several major advances have been made since Baum [123] proposed the original idea of HMM. Baum's work allows estimation of parameters associated with a discrete HMM (i.e., a model in which the probability distribution of observations in each Markov state is discrete) or a continuous density HMM in which the observation density in a state satisfies a log-concavity assumption. This is a serious limitation on this otherwise powerful modeling technique because the more the chosen form of the distribution deviates from that of the true distribution, the less likely it is to be able to achieve Bayes' optimal performance. In 1982, Liporace [124] broadened the class of HMMs that can be estimated by the re-estimation algorithm to elliptically symmetric densities. In 1984, Juang [125] (and subsequently Juang, Levinson, and Sondhi [126]) was successful in eliminating these prior assumptions and limitations on the form of the distribution and showed a method for estimating HMMs with mixture densities (which allow arbitrarily close approximation to the true data distribution). This advance gave HMM a firm foundation for use as a probability distribution of speech for statistical-recognition system designs. Mixture-density HMM has since become the prevalent speech-modeling method and is being used in most speech-recognition systems.

The Search Problem

Hidden Markov models are finite-state automata in nature and form a powerful union when combined with finite-state networks to represent a language (from phonemes to words to grammars that specify the word sequence relationship), particularly for large-vocabulary continuous-speech-recognition systems [127, 128]. Such networks are often very large, and it becomes important to find efficient search methods that evaluate the likelihood that a "path" in such a vast network produces the observed acoustic signal and then find the best among all possible paths.

In the early development of speech recognition, dynamic programming (DP) techniques [107-109] were the focus of the efforts (discussed earlier). Along with the development of the HMM, the fundamental DP technique is now often called the Viterbi algorithm [89].

To deal with large-vocabulary, continuous-speech-recognition problems, the techniques often used are beam search [129], which prunes unlikely events from the search list to achieve efficiency, and the stack algorithm [130], which attempts to find the best path first. New algorithms such as the tree-trellis algorithm [131] which combines a Viterbi forward search and an A^* [132] backward search are very efficient in generating N-best results.

From Bayes to Neyman-Pearson

Bayes' formulation of the pattern-recognition problem assumes that each unknown observation belongs to one of M classes. The maximum a posteriori (MAP) decision rule guarantees optimal performance, i.e., minimum Bayes risk or error, if the joint distribution of the observation and the class, $P(X,C_i)$, is known. In many speech-recognition applications, however, the speech pattern to be recognized may not belong to any of the registered classes. This may appear in the form of so-called "out-of-vocabulary" (OOV) words or as a result of disfluency such as repair or partially spoken words. Another example occurs in a particular telecommunication call-routing application [120] in which the speaker is allowed to embed "keywords" ("collect," "person-to-person," "operator," "credit card," etc.) in naturally spoken sentences (e.g., "I'd like to make a collect call."). In such cases the recognizer needs to be able to distinguish keywords from nonkeywords as well as to identify which keyword has been spoken. For this kind of task, namely detection of target event, a formulation based on hypothesis testing becomes necessary [133, 134].

Let us denote the target event (e.g., a keyword) by E and the nontarget event by $\bar{E}$. The likelihood ratio test performed on an unknown observation, X, is defined as

$$\frac{P(X|E)}{P(X|E)}\begin{cases}\geq \tau, & \text{then } X \in E\\ < \tau, & \text{then } X \in E\end{cases}$$

The likelihood ratio is an important parameter for the calculation of a confidence measure. The threshold defines an operating point on the ROC (receiver operating characteristic) curve for a desired tradeoff between misdetection and false-alarm (false-triggering) errors. For many voice command and control applications, the ability to avoid false triggering by spurious sounds is critically important.

The Neyman–Pearson formalism is also the basis of a new approach to speech understanding focusing on key words and key phrases that carry the main intention or meaning that the speaker would like to deliver [135].

Language Modeling

Just as the goal of acoustic modeling is to find the regularities and variability in the realization of words and phrases, the aim of a language model is to find and represent the relationship among words in sentences. Traditionally, word relationships are expressed in terms of a grammar (e.g., [136]). Shannon's information theory spawned a new perspective in language modeling in which word sequence relationships are expressed as conditional probabilities. If W is a sequence of words:

$$W = w_1 w_2 \cdots w_Q$$

then

$$\begin{aligned} P(W) &= P\left(w_1 w_2 \cdots w_Q\right) \\ &= P(w_1)P\left(w_2|w_1\right)P\left(w_3|w_2 w_1\right) \\ &\quad \cdots P\left(w_Q|w_{Q-1}\cdots w_1\right) \end{aligned}$$

The ensemble of the conditional probabilities (often truncated to length N, $P(w_Q|\, w_{Q-1} \ldots w_{Q\text{-}N+1})$, the so-called N-gram) forms a probabilistic model of the language. Specific values of the conditional probabilities can be estimated from a large text data set via methods such as maximum likelihood training [104]. This corresponds to modeling a language with a finite-state stochastic grammar that can be effectively used in practice. Although such a grammar often overgenerates with respect to a natural language grammar, it has the advantage of complete coverage of natural sentences. If the model is well trained, then ungrammatical sentences would have lower probabilities than the grammatical ones.

The statistical language model has been shown to be effective in large-vocabulary speech recognition. However, its interaction with the acoustic model, in terms of the overall accuracy for speech-to-text conversion, is still not well understood. A language model that achieves lower perplexity (average word-branching factor) for a particular (text) database may not necessarily lead to higher recognition accuracy.

The Robustness Problem

The statistical approach to speech recognition relies heavily on the training data that is available for creating the reference models. The closer the collected training data is to the actual signal encountered during operation, the higher the recognition accuracy is expected to be. The variability in speech, however, comes from many factors and is so large (and at times heterogeneous) that only in rare cases is the amount of collected speech data considered to be reliably sufficient. What is often observed in speech-recognition applications is that a recognizer designed on a data set in the laboratory does not perform as well in the field. In other words, a mismatch between the modeling (training) and the operating (testing) conditions usually exists and causes degradation in the recognizer's performance [137].

Besides the mismatch, several adverse conditions are also often present during operation, such as ambient and transmission noise, distortions due to room acoustics and transducers, and even changes in speech characteristics due to psychological awareness of talking to a machine [137]. These conditions need to be dealt with in order for the recognizer to be able to deliver reliable results. This is the so-called "robustness" problem in automatic speech recognition.

One method that achieves robust results is to collect an extremely large amount of data that reflects the actual operating conditions of the recognizer. With a proper data set, multi-style training [138] was shown to be effective. When the distortion is mostly linear, cepstral compensation in the form of cepstral mean subtraction [139] and cepstral bias removal [140] is simple and works well. More recent advances in robust speech recognition include parallel model combination [141], maximum a posteriori adaptation [142, 143], and stochastic matching [144].

In spite of these developments, the robustness problem remains today an active research area in speech recognition.

Other Advances

While the paradigm shift to statistical methods put speech-recognition research on a mathematically sound basis, it also exposed the limitation of our knowledge in pursuing the Bayes minimum error. Recall that the optimal performance of a recognition system, in terms of the error rate, is attainable only when complete, accurate knowledge on the joint observation-class distribution is available to the designer. Practically, the distribution can only be approximated and, therefore, the distribution estimation approach cannot guarantee any optimality. To circumvent this problem, in order to obtain best accuracy given the choice (form) of the recognizer structure (or distribution function), the method of minimum classification error with a generalized probabilistic descent algorithm [122, 145] was shown to be extremely effective and suitable for speech-recognition applications. This process is known as discriminative training.

Another important methodological advance is adaptive training. Adaptation of system parameters is necessary in the following scenarios:

▲ A speaker-dependent system trained on a particular speaker is to be used by another speaker;

▲ A speaker-independent system needs to deliver improved performance for a specific speaker;

▲ A system needs to adapt to the operating environment to deliver high, robust, performance; or

▲ A speaker-dependent system needs to track changes in the speaker's speech characteristics (e.g., as a result of catching a cold).

The maximum *a posteriori* formulation has been proposed as a framework [142, 143]. This is also an active research area at present.

Spoken-Language Understanding

Except for dictation and some simple command and control applications, speech recognition (transcribing the words spoken) is not nearly as useful as speech understanding (interpreting those words). Although spoken language has been used for centuries by humans to interactively solve problems, it is only in recent years that it has begun to be used in human-machine interfaces. It is also only in recent years that it is possible to envision technology that makes speech as accessible as text as an information source. This section outlines progress in spoken-language understanding over the past 50 years, summarizes current applications in database query and information extraction, and discusses future possibilities.

A Brief History

Spoken-language understanding as undertaken at present involves integrating speech recognition (what are the words?) and natural language understanding (what do those words mean?). The past 50 years have witnessed dramatic changes in each of these component technologies. Some of these changes in speech recognition have already been reviewed in this article. Dramatic changes have also taken place in language understanding. Two important books crystalizing a formal approach to language appeared in 1951, one more influenced by algebra (Zellig Harris's *Methods in Structural Linguistics* appeared in 1951), and one more influenced by psychology and the processing of information by humans (George Miller's *Language and Communication*). Taken together, these works made it possible to imagine the possibility of automatic speech understanding as the computation of an abstract representation and extraction of information.

In the late 1950s and early 1960s, one of Harris's students, Noam Chomsky, promoted a new view of the proper study of linguistics. This view built on the formal methods developed by Harris but replaced the previous focus on language analysis with a new focus on language *generation*. This work was influential in advances in speech synthesis, and it could have served as an important complement to the earlier analytical work (since, normally, people both*generate* and *understand* language). However, the impact was to define linguistics for a large share of language researchers as the study of how to generate speech from the "perfect" speaker-hearer. This dramatically limited the usefulness of linguistics in language understanding since analysis (not just generation) is required for understanding, and since understanding of "imperfect" input needs to be accounted for. A side-effect was the interpretation of "data": instead of being what people actually said, data came to be interpreted as the linguist's intuitions about what the ideal speaker would say. Such methods and goals that are so different from those of engineers led to somewhat of a cultural gap between "linguistic knowledge" and "speech knowledge." Successful speech understanding requires the bridging of this gap.

In the 1960s and 1970s, as socio-linguists and anthropological linguists remained focussed on observing actually occurring language, computational linguists began linguistically relevant computations. However, it was only about 10 years ago that the natural-language-understanding community began to change the trend from the use of "typical" examples based on intuitions to test their systems to the use of data from humans producing language in a communicative setting.

Efforts over the last 10 years show an increasing impact of the two fields on each other (see, e.g., [146]-[151]). Although the use of linguistic knowledge and techniques in engineering may have lagged the use of statistical methods in computational linguistics, there are signs of growth in this area as engineers tackle the more abstract linguistic units (with and without collaboration with natural language experts). These more-abstract units are more rare, and therefore more difficult to model by standard, data-hungry engineering techniques. However, perhaps the biggest recent development for both speech and language understanding has been the use of more realistic data. This focus, partially driven by funding sources (e.g., DARPA) in search of more near-term applications, has led to some basic research toward theories that can accommodate the broadest class of language use: we will be able to "generalize" more of what we learn from working on conversational speech to recognizing isolated digits than we would be able to generalize from digits to conversational speech.

Present Focus: Database Query and Information Extraction

Natural-language understanding presently focuses on applications of the following two classes: database query systems and information extraction systems. A natural-language-database query system formulates a query, usually based on one or a few sentences, into a specification of information fields and values in the context of the particular database's structure. An information-extraction system aims at detecting or summarizing information of interest from a report (e.g., a newswire story or broadcast) in general domains. A pioneering effort that utilized formal models of linguistic structure for "database query" was the work by Levinson and Shipley [152], which preceded much of the current focus and taxonomy of approaches.

Evaluation of spoken-language-understanding systems is required to estimate the state of the art objectively. However, evaluation itself has been one of the challenges of spoken-language understanding. The only systematic program with broad participation for assessing speech understanding has been the (D)ARPA benchmarks focussed on the air travel planning domain (see [153]-[155]). Since it has not yet been possible to agree on a representation for meaning, these evaluations were carried out by human assessment of the results of a database query. Trained annotators translated the human queries into formal database queries with additional annotations for ambiguities and context dependencies. For example, a query, "I want flights from Boston to DC" is expected to produce a table of flights, listing carriers, flight numbers, departing times, and arrival times, etc. Annotation of this type proved to be an expensive proposition, and yet it did not allow for the evaluation of the interactive aspect of the task, since systems were evaluated only on the results returned from a database. Although one test set was set aside for future evaluations, these tests have not been used since 1994.

In the last Air Travel Information Service (ATIS) evaluation of DARPA (December 1994) [155] the speech-recognition word error rate in the best system was under 2%; utterance error rates were about 13% to 25%. The utterance-understanding error rates ranged from 6% to 41%, although about 25% of the utterances were considered impossible to evaluate in the testing paradigm (the trained annotator could not determine what the correct response should be). Hence, these figures do not consider quite the same set. For limited domains, these error rates are probably adequate for many potential applications. Since conversational repairs in human-human dialogue can often be in the ranges observed for these systems, the bounding factor in applications may be not the error rates so much as the ability of the system to manage and recover from errors.

The state of the art in information extraction, based on the DARPA Message Understanding Conference (MUC) evaluations, spans a wide range [156]. Information extraction addresses the problem of updating structured databases (relational or object-oriented) from speech or text. For instance, suppose one needs to update a database of the officers of a corporation with the positions that they hold from broadcast news or newswire stories that report changes in the company officers. The goal is not only to scan source speech and text for such announcements, but also to automatically update the database. For the "named entity" application, where the system has to find all named organizations, locations, persons, dates, times, monetary amounts, and percentages, the error rate from text is 5%. For the "scenario template" application, where the system has to extract complex (but prespecified) relationships in well-defined domains (such as changes in corporate officers) in an open source (such as the Wall Street Journal), the error rate for finding the correct elements of the templates is around 45%.

Researchers are still discussing possibilities for some type of limited speech understanding that would be less costly and more relevant in applications. Part-of-speech tagging has been discussed, but it has not been shown that good part-of-speech tagging is either necessary or sufficient for good understanding. Other possibilities include dividing spoken conversations into linguistic units more like sentence and phrase boundaries, finding the main verb (if any) in that unit, and/or indicating words with extra emphasis.

Future Challenges

Speech-understanding research was nonexistent 50 years ago. The dramatic changes in speech recognition and in language understanding during the past 50 years, combined with political changes and changes in the computing infrastructure, led to the state of the art that we observe today. Challenges remain in several areas (see [157]):

▲ **Integration.** There is much evidence that human speech understanding involves the integration of a great variety of knowledge sources, including knowledge of the world or context, knowledge of the speaker and/or topic, lexical frequency, previous uses of a word or a semantically related topic, facial expressions (in face-to-face communication), prosody, in addition to the acoustic attributes of the words. Our systems could do much better by integrating these knowledge sources.

▲ **Prosody.** Prosody can be defined as information in speech that is not localized to a specific sound segment, or information that does not change the identity of speech segments (see, e.g., [158], [159], [160]). Such information includes the pitch, duration, energy, stress, and other supra-segmental attributes. The segmentation (or grouping) function of prosody may be related more to syntax (with some relation to semantics), while the saliency or prominence function may play a larger role in semantics and pragmatics than in syntax. To make maximum use of the potential of prosody will likely require a well-integrated system, since prosody is related to linguistic units not just at and below the word level, but also to abstract units in syntax, semantics, discourse, and pragmatics. Our systems make quite limited (or no) use of prosody at present.

▲ **Spontaneous Speech**. The same acoustic attributes that indicate much of the prosodic structure (e.g., pitch, stress, and duration patterns) are also very common in aspects of spontaneous speech that seem to be more related to the speech planning process than to the structure of the utterance. For example, a long syllable followed by a pause can indicate either an important syntactic boundary or that the speaker is planning the rest of the utterance. Similarly, a prominent syllable may mark new or important information, or a restart intended to replace something said in error. Although spontaneous speech effects are quite common in human communication and may be expected to increase in human machine discourse as people become more comfortable conversing with machines, modeling of speech disfluencies is only just beginning (see, e.g., [161], [162]).

Much of our thinking about spoken language has been focused on its use as an interface in human-machine interactions mostly for information access and extraction. With increases in cellular phone use and dependence on networked information resources, and as rapid access to information becomes an increasingly important economic factor, telephone access to data and telephone transactions will no doubt rise dramatically. There is a growing interest, however, in viewing spoken language not just as a means to access information, but as, itself, a source of information. Important attributes that would make spoken language more useful in this respect include: random access, sorting (e.g., by speaker, by topic, by urgency), scanning, and editing. How could our lives be changed by such tools? Enabling such a vision challenges our systems still further in noise robustness and in spontaneous speech effects. Further, the resulting increased accessibility to information from conversational speech will likely also raise increased concern for privacy and security, some of which may be addressed by controlling access by speech: speaker identification and verification (see the next section).

While such near-term application possibilities are exciting, we can envision an even greater information revolution on par with the development of writing systems if we can successfully meet the challenges of spoken language both as a medium for information access *and* as itself a source of information. Spoken language is still the means of communication used first and foremost by humans, and only a small percentage of human communication is written. Automatic-spoken-language understanding can add many of the advantages normally associated only with text (random access, sorting, and access at different times and places) to the many benefits of spoken language. Making this vision a reality will require significant advances.

Speaker Verification and Identification

Speaker recognition is the process of automatically recognizing a speaker by using speaker-specific information included in his or her speech [163-166]. This technique can be used to verify the identity claimed by people accessing systems; that is, it enables control of access to various services by voice. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers.

Speaker recognition can be classified into speaker identification and speaker verification. Closed-set speaker identification is the process of determining which of the registered speakers a given utterance comes from. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Most of the applications in which voice is used to confirm the identity claim of a speaker require speaker verification.

Speaker-recognition methods can also be divided into text-dependent and text-independent methods. The former requires the speaker to provide utterances of key words or sentences that are the same text for both training and recognition, whereas the latter does not rely on a specific, prescribed text. The text-dependent methods are usually based on template-matching techniques in which the time axis of an input speech sample and each reference template or reference model of the registered speakers are aligned, and the similarity between them is accumulated from the beginning to the end of the utterance [164, 167, 168]. Since this method can directly exploit voice individuality associated with each phoneme or syllable, it generally achieves higher-recognition performance than the text-independent model.

However, there are several applications, such as forensic and surveillance applications, in which predetermined keywords cannot be used. Moreover, human beings can often recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have recently attracted more attention. Another advantage of text-independent recognition is that it can be done sequentially, until a desired level of significance is reached, without the annoyance of the speaker having to repeat the key words again and again.

Both text-dependent and text-independent methods have a serious weakness. These systems can easily be defeated, because someone who plays back the recorded voice of a registered speaker uttering key words or sentences into the microphone can be accepted as the registered speaker. To cope with this problem, a text-prompted speaker-recognition method has recently been proposed.

Basic Structures of Speaker-Recognition Systems

The fundamental techniques, such as signal analysis, modeling and pattern matching, in a speaker identification/verification system are essentially identical to those used in a speech-recognition system. What differentiates them is the need to find speaker-specific information and the explicit use of hypothesis analysis and thresholding.

In the closed-set speaker-identification task, a speech utterance from an unknown speaker is analyzed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an identity claim is made by an unknown speaker, and an utterance of this unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is good enough, that is, above a threshold, the identity claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system, but at the price of falsely rejecting valid users. Conversely, a low threshold enables valid users to be accepted consistently, but at the price of accepting impostors. To set the threshold at the desired level of customer rejection and impostor acceptance, it is necessary to know the distribution of customer and impostor scores.

The effectiveness of speaker-verification systems can be evaluated by using the receiver operating characteristics (ROC) curve, which shows the system performance in terms of two probabilities: the probability of correct acceptance and the probability of incorrect acceptance. By varying the decision threshold, a point on the ROC curve can be selected for operating purposes (to achieve the desired tradeoff between the two probabilities) [169]. The equal-error rate (EER) is commonly accepted as an overall measure of the system performance. It corresponds to the threshold at which the false acceptance rate is equal to the false rejection rate.

From Spoken Language to Multimodal Communication

Human-machine communication (HMC) is evolving from text interface (i.e., keyboard and screen display) to spoken language (automatic speech recognition and understanding) to multimodal communication involving different senses (audio, visual, tactile, or even gestural) with synergy [170, 171]. Human communication includes the perception or production of a message or of an action as an explicit or implicit cognitive process. For perception, there are the "five senses": hearing, vision, touch, taste, and smell, with reading as a specific visual operation, and speech perception as a specific hearing operation. For production, it includes sound (speech, or general sound production) and vision (generation of drawings, graphics or, more typically, written messages). Cognition includes the means to understand or to generate a message or an action from a knowledge source.

The machine serves as a means for the human being to communicate with the world. In the domain of HMC, the computer has various artificial perception abilities: speech, character, graphics, and gesture or movement recognition. This recognition function can be accompanied by the recognition of the identity of the person through the same modes. Gesture or movement recognition is made through the use of special equipment (such as the VPL DataGlove or DataSuit, or the Cyberglove), which includes position sensors. Other sensors allow for recognizing the direction of viewing (through an oculometer or through a camera). Reciprocally, the computer can produce messages using various modes ranging from the display of a textual or graphical (including icons) message to concept-to-text generation or summary generation, speech synthesis, and static or animated image synthesis. The visual information can be produced in stereovision or within a complete environment in which the user is immersed ("virtual" reality), or it can be superimposed on the real environment ("augmented" reality), which would require the wearing of special equipment. The provided information can be multimedia, including text, real or synthetic images, and sound. It is also possible, in the gestural communication mode, to produce a kinesthetic feedback, allowing for the generation of simulated solid objects.

The machine also needs to have cognitive abilities. It must take into account a model of the user, of the world on which he acts, of the relationship between those two elements, but also of the task that has to be carried out and of the structures of the dialogue. It must be able to reason, to plan a linguistic or nonlinguistic act in order to reach a target, to solve problems and aid in decision making, to merge information coming from various sensors, and to learn new knowledge or new structures. Multimodal communication raises the problem of co-reference (e.g., when the user designates an object, or a spot, on the computer display and pronounces a sentence relative to an action on that object).

To accomplish the goal of multimodal human-machine communication, while it is important to understand the human functions in order to get some inspiration when designing a system, of greater importance is the ability to model in the machine the user with whom it has to communicate. It is also necessary to model the world in which they occur. This extends HMC to various research domains such as room acoustics, physics, or optics, and also physiology and cognitive psychology (for generating *intelligent agents* or avatars).

Linking Language and Image

With the coming of "intelligent" images, the relationship between language and image is getting closer [172]. It justifies advanced human-machine communication modes. In an "intelligent" synthetic image (which implies the modeling of physical characteristics of the real world), a sentence such as "Throw the ball on the table" will induce a complex scenario where the ball will rebound on the table, then fall on the ground. This scenario would be difficult to describe to the machine with usual low-level computer languages or interfaces. Visual communication is directly involved in human-to-machine communication (e.g., for recognizing the user or the expressions on his face), but also indirectly involved in the building of a visual reference that will be shared by the human and the machine, allowing for a common understanding of the messages that they exchange (for example, in the understanding of the command "Take the knife which is on the small marble table" addressed to a robot). Instead of considering the user on one side and the machine on the other side, the user himself may become an element of the simulated world: acting and moving in this world, and getting reactions from it.

There are several similarities in the research concerning these different communication modes. In speech, vision, and gesture processing, similar methods are used for signal processing, coding and pattern recognition. The same approach based on statistical modeling has been applied with similar algorithms to various domains of HMC such as speech recognition, visual recognition of character or object, or gesture [173]. This approach requires large databases, which are now available for speech, characters and text data, but have yet to be made available for visual, gestural, and multimodal data.

Humans use multimodal communication when they speak to each other, except in the case of pathology or of telephone communication. Movements of the face and lips, as well as expression and posture, will be involved in the spoken language communication process. Studies in speech intelligibility also showed that having both visual and audio information improves the information communication, especially when the message is complex or when the communication takes place in a noisy environment [174], [175]. This has led to studies in bimodal speech synthesis and recognition.

In the field of speech synthesis, models of speaking faces were designed and used in speech dialogue systems [176]. The face and lip movements were synthesized by studying those movements in human speech production through image analysis. It resulted in text-to-talking heads synthesis systems. Studies in using the visual information in speech communication (e.g., using the image of the lips only, or the bottom of the face or the entire face) showed that the intelligibility of the synthesized speech was improved for the human "listener," especially in a noisy environment. In the same way, the use of the visual face information, and especially the lips, in speech recognition was studied, and results showed that using both types of information gives better recognition performances than using only the audio or only the visual information, especially in a noisy environment [177, 178].

While this visual information on the human image can be used as part of the spoken-language-communication process, other types of visual information related to the human user can also be considered by the machine. The fact that the user is in the room, or is seated in front of the computer display, as well as the direction of his/her gaze can be used in the communication process (e.g., waiting for the presence of the human in the room to synthesize a message, or choosing between a graphic or spoken mode for delivering information, depending of whether the user is in front of the computer or somewhere else in the room, adjusting the synthesis volume depending on how far he is from the loudspeaker, adapting a microphone array on the basis of the position of the user in the room [179], checking what the user is looking at on the screen in order to deliver information relative to that area [180], etc.)

Multimodal Multimedia Communication

Communication can also involve several verbal and nonverbal media. Berkley and Flanagan [181] designed the AT&T Bell Labs HuMaNet system for multipoint conferencing over the public telephone network. The system features hands-free sound pick up through microphone arrays, voice control of call set-up, data access and display through speech recognition, speech synthesis, speaker verification for privileged data, still image and stereo image coding. It has been extended to also include tactile interaction, gesturing and handwriting inputs, and face recognition [182]. In Japan, ATR has a similar advanced teleconferencing program, including 3D object modeling, face modeling, voice command, and gestural communication. At IRST, Stringa et al. [183] have designed, within the MAIA project, a multimodal interface (speech recognition and synthesis, and vision) to communicate with a "concierge" of the institute, which answers questions on the institute and its researchers, and with a mobile robot, which has the task of delivering books or accompanying visitors.

In the ESPRIT "Multimodal-Multimedia Automated Service Kiosk" (MASK) project, speech recognition and synthesis are used in parallel with other input (touch screen) and output (graphics) means [184]. The application is to provide railway travel information to railway customers, including the possibility of making reservations. The users get both visual (graphics) and audio (speech synthesis) information, and they may choose to either use speech or tactile input. First studies show that subjects tend to use one mode or the other, based on its apparent reliability or on their own preference, but they will not mix them up during the dialog.

In the closely related domain of multimedia information processing, interesting results have been obtained in the Informedia project at CMU on the automatic indexing of TV broadcast data (news), and multimedia information query by voice. The system uses continuous speech recognition to transcribe the speech. It segments the video information in sequences, and uses natural-language-processing techniques to automatically index those sequences from the result of the textual transcriptions. Although the speech recognition is far from being perfect (about 50% recognition rate), it seems to be good enough for allowing the user to get a sufficient amount of multimedia information from his queries [185].

Current Speech Processing Technical Committee Members

Alejandro Acero, *Microsoft*
Jean-Pierre Adoul, *University of Sherbrooke*
Abeer Alwan, *UCLA*
Juin-Hwey Chen, *Voxware*
Li Deng, *University of Waterloo*
Ponani Gopalakrishnan, *IBM*
John D. Lafferty, *Carnegie Mellon University*
Lori Lamel, *LIMSI-CNRS*
Chin-Hui Lee, *Lucent Technologies-Bell Labs (Chair)*
Matthew Lennig, *Nuance Communications*
Alan McCree, *Texas Instruments*
Takehiro Moriya, *NTT*
Eric F. Moulines, *Ecole Nationale Superieure des Telecommunications*
Hermann Ney, *RWTH Aachen - University of Technology*
Yves Normandin, *Locus Speech Corp.*
Raja Rajasekaran, *Texas Instruments*
Douglas Reynolds, *Lincoln Laboratory*
J. Robin Rohlicek, *Fish and Richardson P.C.*
Shigeki Sagayama, *NTT, Japan*
Steve Young, *Cambridge University*
Marc A. Zissman, *Lincoln Laboratory*

Conclusion

We attempted to provide a comprehensive, albeit cursory, review of how speech signal-processing technologies progressed in the past as well as the challenges ahead. Speech processing is one of the most intriguing areas of intelligent signal processing because humans generate, use, and appreciate speech on a daily basis. Speech research has attracted scientists as an important discipline and has created technological impact on society and is expected to further flourish in this era of machine intelligence and human-machine interaction. We hope this article brings about understanding as well as inspiration.

Acknowledgments

In the course of writing this article, we received valuable comments and suggestions from Peter Kroon, Steve Levinson, Larry Rabiner, Raj Reddy, C.H. Lee, K. Lenzo, P. Placeway, Kristie Seymore, Mathew Siegler, and Victor Zue. Special thanks are due to them for their generosity and efforts.

References

1. W.V. Kempelen. *Le mechanisme de la parole, suivi de la description d'une machine parlante*. J.V. Degen, Vienna, 1791.

2. Sir Charles Wheatstone. *The Scientific Papers of Sir Charles Wheatstone*. Taylor & Francis, London, 1879.

3. H. Dudley. Synthesis speech. *Bell Labs. Record*, 15:98-102, 1936.

4. J.L. Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin, Germany, 1972.

5. D. Childers *Spectral Estimation*. IEEE Press.

6. G. Fant. *Acoustic Theory of Speech Production*. Mouton and Co., XX's-Gravenhage, The Netherlands, 1960.

7. N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper & Row Publishers, New York, 1968.

8. J.L. Flanagan. Note on the design of terminal analog speech synthesizers. *J. Acoust. Soc. Am.*, 29:306-310, 1957.

9. B.S. Atal and S.L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2):637-655, 1971.

10. F. Itakura and S. Saito. Analysis synthesis telephony based upon the maximum likelihood method. In Y. Kohasi, editor, *Reports of 6th Int. Cong. Acoust.*, pages C-5-5, C-17-20, Tokyo, 1968.

11. H. Wakita. Direct estimation of vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Trans.*, AU-21:417-427, 1973.

12. M.M. Sondhi. New methods of pitch extraction. *IEEE Trans. Audio and Electroacoustics*, AU-16(2):262-266, June 1968.

13. M.R. Schroeder. Period histogram and product spectrum: New methods for fundamental frequency measurement. *J. Acoust. Soc. Am.*, 43(4):829-834, April 1968.

14. A.M. Noll. Pitch determination of human speech by the harmonic product spectrum: The harmonic sum spectrum and a maximum likelihood estimate. *Proc. Symp. Computer Proc. in Comm.*, pages 779-798, April 1969.

15. L. Rabiner, B. Juang, S. Levinson, and M. Sondhi. Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition. *In Proc. of IEEE ICASSP-85*, pages 9-12, Tampa, Florida, 1985.

16. S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. ASSP, 34(1):52-59, 1986.

17. D.H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, 82:737-793, 1987.

18. J. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg. *Progress in Speech Synthesis*. Springer, New York, 1997.

19. C.H. Coker. Synthesis by rule from articulatory parameters. *In Proc. 1967 Conf. Speech Comm. Process.*, pages 52-53, Boston, MA, 1967.

20. J.L. Flanagan and L. Cherry. Excitation of vocal-tract synthesizers. *J. Acoust. Soc. Am.*, 45(3):764-769, 1968.

21. J.K. Kelly, Jr. and C.C. Lochbaum. Speech synthesis. *Proc. Fourth Intern. Congr. Acoust.*, G42:1-4, 1962.

22. S. Maeda. On a simulation method of dynamically varying vocal tract reconsideration of the Kelly-Lochbaum Model. In R. Carre, R. Descout, and M. Wajskop, editors, *Articulatory Modeling and Phonetics*, pages 281-288. Groupe de Communication Parlee, Grenoble, France, 1977.

23. P. Rubin, T. Baer, and P. Mermelstein. An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.*, 70(2):321-328, 1981.

24. H.W. Strube. Time-varying wave digital filters and vocal-tract models. *Proc. ICASSP*, pages 923-926, 1982.

25. J.L. Flanagan and I.L. Landgraf. Self-oscillating source for vocal tract synthesizers. *IEEE Trans. Audio and Electroacoustics*, 16:57-64, 1968.

26. J.L. Flanagan and K.L. Ishizaka. Automatic generation of voiceless excitation to a vocal cord-vocal tract speech synthesizer. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(2):163-170, 1976.

27. J.L. Flanagan and K. Ishizaka. Computer model to characterize the air volume displaced by the vibrating vocal cords. *J. Acoust. Soc. Am.*, 63:1559-1565, 1978.

28. E. L. Bocchieri and D. G. Childers. An animated interactive graphics editor for the study of speech articulation. *Speech Technology*, 2:10-14, 1984.

29. D.G. Childers and C. Ding. Articulatory synthesis: Nasal sounds and male and female voices. *J. Phonet.*, 19:453-464, 1991.

30. D.R. Allen and W.J. Strong. A model for the synthesis of natural sounding vowels. *J. Acoust. Soc. Am.*, 78(1):58-69, 1985.

31. M.M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. Acoust., Speech, and Signal Processing*, 35(7):955-967, 1987.

32. A. Fettweis and K. Meerkotter. On adaptors for wave digital filters. *IEEE Trans. Acoustics, Speech, Signal Proc.*, 23(6):516-525, 1975.

33. P. Meyer and H.W. Strube. Calculations on the time-varying vocal tract. *Speech Comm.*, 3:109-122, 1984.

34. B. Gold and L.R. Rabiner. Analysis of digital and analog formant synthesizers. *Trans. Audio Electroacoustics*, AU-16:81-94, 1968.

35. J.N. Holmes. Formant synthesizers: cascade or parallel. *Speech Commun.*, 2(4):251-274, 1983.

36. D.H. Klatt. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, 67(3):971-995, 1980.

37. N.B. Pinto, D.G. Childers, and A.L. Lalwani. Formant speech synthesis: improving production quality. *IEEE Trans. Acoust., Speech, Signal Process.*, 37(12):1870-1887, 1989.

38. L. Rabiner and C. Rader. *Digital Signal Processing*. IEEE Press, New York, 1972.

39. L. Rabiner and B. Gold. *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

40. J. Markel and A.H. Gray. *Linear Prediction of Speech*. Springer-Verlag, New York, 1980.

41. B. Hanson, B. Juang, and D. Wong. Speech enhancement with harmonic synthesis. *InProc. of ICASSP-83*, pages 1122-25, Boston, 1983.

42. R.J. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-34:744-754, 1986.

43 E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.*, 9:453-467, 1990.

44. N.S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

45. W.R. Bennett. Spectra of quantized signals. *Bell System Tech. J.*, pages 446-472, July 1948.

46. A. Gersho. Principles of quantization. *IEEE Transactions on Circuits and Systems*, pages 427-436, 1978.

47. H. Gish and J.N. Pierce. Asymptotically efficient quantizing. *IEEE Trans. on Information Theory*, pages 676-683, September 1968.

48. B. Smith. Instantaneous companding of quantized signals. *Bell System Tech. J.*, pages 653-709, May 1957.

49. N.S. Jayant. Adaptive quantization with a one word memory. *Bell System Tech. J.*, pages 1119-1144, September 1973.

50. D.J. Goodman and A. Gersho. Theory of an adaptive quantizer. *IEEE Trans. on Communications*, pages 1037-1045, August 1974.

51. C.B. Rubinstein and J.O. Limb. On the design of quantizers for DPCM coders: The influence of subjective testing methodology. *IEEE Trans. on Communications*, pages 565-573, May 1978.

52. R.W. Stroh. Differential PCM with adaptive quantization for voice communications. *Proc. Int. Conf. Communications*, pages 130.1-130.5, June 1974.

53. D.W. Petr. 32 kb/s ADPCM-DLQ coding for network applications. *In Proc. IEEE Global Telecomm. Conf.*, pages A8.3-1-A8.3-5, 1982.

54. B. Atal and M.R. Schroeder. Predictive coding of speech signals. *Bell System Tech. J.*, pages 1973-1986, October 1970.

55. B.S. Atal and M.R. Schroeder. Predictive coding of speech signals and subjective error criteria. *IEEE Trans. Speech Signal Proc.*, ASSP-27(3):247-254, 1979.

56. B.S. Atal and M.R. Schroeder. Stochastic coding of speech at very low bit rates. *InProc. Int. Conf. Comm.*, pages 1610-1613, Amsterdam, 1984.

57. J. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE. J. Sel. Area Comm.*, 6(2), 1988.

58. D. Griffin and J.S. Lim. Multiband excitation vocoder. *IEEE Trans. ASSP*, 36(8):1223-1235, 1988.

59. L. Supplee, R. Cohn, J. Collura, and A. McCree. MELP: The new Federal Standard at 2400 bps. *In Proc. ICASSP-1997*, pages 1591-1594, Munich, Germany, 1997.

60. W.B. Kleijn and J. Haagen. Transformation and decomposition of the speech signal for coding. *IEEE Signal Process. Lett.*, 1(9):136-138, 1994.

61. B. Gold and C. Rader. Systems for compressing the bandwidth of speech. *IEEE Trans. Audio & Electroacoust.*, AU-15(3), 1967.

62. J.D. Markel and A.H. Gray, Jr. A linear prediction vocoder simulation based upon the autocorrelation method. *IEEE Trans. ASSP*, 22:124-134, 1974.

63. A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publisher, Dordrecht, Holland, 1991.

64. A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel. Speech coding based upon vector quantization. *IEEE Trans. ASSP*, 28:562-574, October 1980.

65. D.Y. Wong, B.H. Juang, and A.H. Gray, Jr. An 800 bit/s vector quantization LPC vocoder. *IEEE Trans. ASSP*, 30:770-780, October 1982.

66. H. Abut, R.M. Gray, and G. Rebolledo. Vector quantization of speech and speech-like waveforms. *IEEE Trans. ASSP*, 30:423-435, June 1982.

67. Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans. COM-28*, pages 84-95, January 1980.

68. B.H. Juang and A.H. Gray, Jr. Multiple stage vector quantization for speech coding. *InIEEE ICASSP-82*, pages 597-600, Paris, May 1982.

69. R.V. Cox. Speech coding standards. In Kleijn et al., editor, *Speech Coding and Synthesis*. Elsevier, New York, 1995.

70. W.R. Daumer, X. Maitre, P. Mermelstein, and I. Tokizawa. Overview of the 32 kb/s ADPCM algorithm. *Proc. IEEE Global Telecomm. Conf.*, pages 774-777, 1984.

71. J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant, and M.J. Melchner. A low-delay CELP coder for the CCITT 16 kb/s speech coding standard. *IEEE J. Selected Areas Comm.*, 10(5):830-849, 1992.

72. P. Mermelstein. G.722, A new CCITT coding standard for digital transmission of wideband audio signals. *IEEE Comm. Magazine*, 26(1):8-15, January 1988.

73. P. Vary, R. Hoffman, K. Hellwig, and R. Sluyter. A regular-pulse excited linear predictive code. *Speech Comm.*, 7(2):209-215, 1988.

74. I.A. Gerson and M.A. Jasiuk. Vector sum excited linear prediction (VSELP). In B. S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, pages 69-79. Kluwer Academic Publishers, 1991.

75. P.J.A. DeJaco, W. Gardner, and C. Lee. QCELP: The North American CDMA digital cellular variable rate speech coding standard. In *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pages 5-6, Sainte-Adele, Quebec, 1993.

76. T. Tremain. The government standard linear predictive coding algorithm: LPC-10. *Speech Technology Magazine*, pages 40-49, April 1982.

77. J.P. Campbell, Jr., T.E. Tremain, and V.C. Welch. The Federal standard 1016 4800 bps CELP vocie coder. *Digital Signal Processing*, 1(3):145-155, 1991.

78. B. Edler. Overview on the current development of MPEG-4 audio coding. In *4th International Workshop on Systems, Signals and Image Processing*, Poznan, May 1997.

79. M.R. Schroeder. U.S. Patent #3,180,936, filed 1960, issued April 1965.

80. Sievers and M. M. Sondhi. Unpublished, 1964.

81. Weiss, Aschkenasy, and Parson. In *IEEE Symp. on Speech Recog.*, Pittsburgh, PA, 1974.

82. S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE ASSP*, 27(2):113-120, April 1979.

83. R.J. McAulay and M.L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. ASSP*, 28(2):137-145, April 1980.

84. Y. Ephraim and D. Malah. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Trans. ASSP*, 32:1109-1121, 1984.

85. J.S. Lim and A.V. Oppenheim. Enhancements and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586-1604, December 1979.

86. Y. Ephraim, D. Malah, and B.H. Juang. On the application of hidden Markov models for enhancing noisy speech. *IEEE Trans. ASSP*, 37(12):1846-1856, December 1989.

87. Etter and Moschytz. Noise reduction by noise-adaptive spectral magnitude expansion. *J. Audio Eng.*, pages 341-349, May 1994.

88. E.J. Diethorn. A low-complexity background-noise reduction preprocessor for speech encoders. *In IEEE Workshop on Speech Coding for Telecommunications*, pages 45-46. September 1997.

89. L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliff, New Jersey, 1993.

90. K.H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *J. Acoust. Soc. of America*, 24:637-642, 1952. Cited in S. R. Hyde, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," (originally written as a British Post Office research report in 1968 and appearing as Chapter 11 in Human Communication: A Unified View, David and Denes (editors), McGraw-Hill, 1972.

91. H.F. Olson and H. Belar. Phonetic typewriter. *J. Acoust. Soc. Am.*, 28(6):1072-1081, 1956.

92. D.B. Fry. Theoretical aspects of mechanical speech recognition; and P. Denes, The design and operation of the mechanical speech recognizer at University College London. *J. British Inst. Radio Engr.*, 19(4):211-229, 1959.

93. J.W. Forgie and C.D. Forgie. Results obtained from a vowel recognition computer program. *J. Acoust. Soc. Am.*, 31(11):1480-1489, 1959.

94. J. Suzuki and K. Nakata. Recognition of Japanese vowels—preliminary to the recognition of speech. *J. Radio Res. Lab*, 37(8):193-212, 1961.

95. T. Sakai and S Doshita. The phonetic typewriter, information processing 1962. *InProc. IFIP Congress*, Munich, 1962.

96. K. Nagata, Y. Kato, and S. Chiba. Spoken digit recognizer for Japanese language. *NEC Res. Develop.*, 6, 1963.

97. 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.

98. T.K. Vintsyuk. Speech discrimination by dynamic programming. *Kibernetika*, 4(2):81-88, January-February 1968.

99. D.R. Reddy. An approach to computer speech recognition by direct analysis of the speech wave. Tech. Report C549, Computer Science Dept., Stanford Univ., September 1966.

100. V.M. Velichko and N. G. Zagoruyko. Automatic recognition of 200 words. Int. *J. Man-Machine Studies*, 2:223, June 1970.

101. H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-26(1):43-49, February 1978.

102. F. Itakura. Minimum prediction residual applied to speech recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23(1):67-72, February 1975.

103. C.C. Tappert, N. R. Dixon, A. S. Rabinowitz, and W. D. Chapma. Automatic recognition of continuous speech utilizing dynamic segmentation, dual classification, sequential decoding and error recovery. Tech. Report TR-71-146, Rome Air Dev. Cen, Rome, NY, 1971.

104. F. Jelinek, L. R. Bahl, and R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Information Theory*, IT-21:250-256, 1975.

105. F. Jelinek. The development of an experimental discrete dictation recognizer. *Proc. IEEE*, 73(11):1616-1624, 1985.

106. 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, Signal Proc.*, ASSP-27:336-349, August 1979.

107. H. Sakoe. Two level DP matching—a dynamic programming based pattern matching algorithm for connected word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27:588-595, December 1979.

108. J.S. Bridle and M.D. Brown. Connected word recognition using whole word templates. *Proc. Inst. Acoust. Autumn Conf.*, pages 25-28, November 1979.

109. C.S. Myers and L.R. Rabiner. A level building dynamic time warping algorithm for connected word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-29:284-297, April 1981.

110. C.H. Lee and L.R. Rabiner. A frame synchronous network search algorithm for connected word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, 37(11):1649-1658, November 1989.

111. J. Ferguson, editor. *Hidden Markov Models for Speech*. IDA, Princeton, NJ, 1980.

112. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257-286, February 1989.

113. D. Klatt. Overview of the ARPA speech understanding project. In W. Lea, editor, *Trends in Speech Recognition*, pages 249-271. Prentice-Hall, NJ, 1980.

114. K.F. Lee, H.W. Hon, and D.R. Reddy. An overview of the SPHINX speech recognition system. *IEEE Trans. Acoustics, Speech, Signal Proc.*, 38:600-610, 1990.

115. Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, et al. BBYLOS: The BBN continuous speech recognition system. *Proc. ICASSP 87*, pages 89-92, April 1987.

116. D.B. Paul. The Lincoln robust continuous speech recognizer. In *Proc. ICASSP 89*, pages 449-452, Glasgow, Scotland, May 1989.

117. M. Weintraub et al. Linguistic constraints in hidden Markov model based speech recognition. *InProc. ICASSP 89*, pages 699-702, Glasgow, Scotland, May 1989.

118. V. Zue, J. Glass, M. Phillips, and S. Seneff. The MIT summit speech recognition system: A progress report. *Proc. DARPA Speech and Natural Language Workshop*, pages 179-189, February 1989.

119. C.H. Lee, L.R. Rabiner, R. Pieraccini, and J.G. Wilpon. Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4:127-165, 1990.

120. B.H. Juang, R. Perdue, and D. Thomson. Deployable automatic speech recognition systems: Advances and challenges. *AT&T Technical Journal*, 74(2), 1995.

121. L. Rabiner and B. Juang. An introduction to hidden Markov model. *IEEE ASSP Magazine*, 3(1):4-16, January 1986.

122. B. Juang, W. Chou, and C.H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech & Audio Proc. T-SAP*, 5(3):257-265, May 1997.

123. L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41:164-171, 1970.

124. L.R. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Trans. Inform. Theory*, IT-28:729-34, September 1982.

125. B. Juang. Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Technical Journal*, 64(6):1235-1250, July-August 1985. Part 1.

126. B.H. Juang, S.E. Levinson, and M.M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Trans. Information Theory*, IT-32(2):307-309, March 1986.

127. J.K. Baker. Stochastic modeling for automatic speech understanding. In D.R. Reddy, editor, *Speech Recognition: Invited Papers of the IEEE Symp.* 1975.

128. L. Bahl et al. Automatic recognition of continuously spoken sentences from a finite state grammar. In *Proceedings ICASSP*, Tulsa, OK, 1978.

129. B.T. Lowerre and R. Reddy. The HARPY speech understanding system. In W. Lea, editor, *Trends in Speech Recognition*, pages 340-360. Prentice-Hall, Englewood Cliffs, NJ, 1980.

130. L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. PAMI*, 5(2):179-190, March 1983.

131. F.K. Soong and E.F. Huang. A tree-trellis fast search for finding N-best sentence hypotheses. In *Proc. ICASSP-91*, pages 705-708, Toronto, May 1991.

132. D. Paul. Algorithms for an optimal A^* search and linearizing the search in the stack decoder. In *IEEE ICASSP-91*, pages 693-696, Toronto, Canada, May 1991.

133. T. Kawahara, C-H. Lee, and B-H. Juang. Key-phrase detection and verification for flexible speech understanding. In *Proc. ICSLP-96*, Philadelphia, PA, October 1996.

134. M. Rahim, C-H. Lee, and B-H. Juang. A study on robust utterance verification for connected digits recognition. *J. Acoustical Society of America*, 1997.

135. T. Kawahara, C-H. Lee, and B-H. Juang. Combining key-phrase detection and subword-based verification for flexible speech understanding. In *Proc. of ICASSP-97*, Munich, April 1997.

136. Z. Harris. *Methods in Structural Linguistics*. University of Chicago Press, 1951. Later updated and published as Structural Linguistics, in 1960 and 1974.

137. B. Juang. Speech recognition in adverse environments. *Computer Speech & Language*, 5:275-294, 1991.

138. Y. Chen. Cepstral domain stress compensation for robust speech recognition. In *Proc. of ICASSP-87*, pages 717-720, Dallas, Texas, April 1987.

139. R.M. Stern, A. Acero, F-H. Liu, and Y. Ohshima. Signal processing for robust speech recognition. In *Automatic Speech and Speaker Recognition—Advanced Topics*. Lee, Soong, and Paliwal (eds.), p. 357-384, Kluwer, 1996.

140. M.G. Rahim and B.H. Juang. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Trans. SAP*, 4(1):19-30, January 1996.

141. M.J.F. Gales and S.J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289-307, 1995.

142. C.-H. Lee and J.-L. Gauvain. Bayesian adaptive learning and MAP estimation of HMM. In C.-H. Lee, F. K. Soong, and K. K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 4. Kluwer Academic Publishers, 1996.

143. B.H. Juang, C.H. Lee, and C.H. Lin. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. Acoustics, Speech, Signal Proc.*, 39(4):806-814, April 1991.

144. A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. on Audio and Speech Processing*, 4(3):190-202, 1996.

145. B-H. Juang and S. Katagiri. Discriminative learning for minimum error training. *IEEE Trans. Signal Processing*, 40(12):3043-3054, December 1992.

146. R. Bobrow and B. Webber. Knowledge representation for syntactic/semantic processing. In *Proc. National Conference on Artificial Intelligence*, pages 99-107, Washington DC, 1980.

147. P. Brown, S. Della Pietra, et al. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311, 1993.

148. B.C. Bruce. Case systems for natural language. *Artificial Intelligence*, 6:327-360, 1975.

149. A. Corazza, R. De Mori, R. Gretter, and G. Satta. Computation of probabilities for an island-driven parser. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):936-950, 1991.

150. R. Kuhn and R. De Mori. The application of semantic classification trees to natural language understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17:449-460, 1995.

151. R. Pieraccini and E. Levin. A learning approach to natural language understanding. In A.J. Rubio Ayuso and J.M. Lopez Soler, editors, *Speech Recognition and Coding. New Advances and Trends.* Springer, Berlin, 1995.

152. S.E. Levinson and K.L. Shipley. A conversational mode airline information and reservation system using speech input and output. *Bell Sys. Tech. Journal*, 59(1):119-137, 1980.

153. D. Pallett. DARPA Resource Management and ATIS Benchmark Test. In *Defense Advanced Research Projects Agency (DARPA) Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, pages 49-58, Pacific Grove, California, 1991.

154. D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Prysbocki. 1993 Benchmark tests for the ARPA spoken language program. In *Proc. Human Language Technology Workshop*, pages 49-74. Morgan Kaufmann, 1994.

155. D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, et al. 1994 Benchmark tests for the ARPA spoken language program. In *Proc. Human Language Technology Workshop*, pages 5-36. Morgan Kaufmann, 1995.

156. DARPA. *Proc. Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland. Morgan Kaufmann Publishers, San Francisco, CA, 1995.

157. P. Price. Spoken language understanding. In *Section 1.8 in Survey of the State of the Art in Human Language Technology*. 1996. R. Cole, editor in chief (http://www.cse.ogi.edu/CSLU/HLTsurvey/ch1node10.html#SECTION18), also, Cambridge University Press, 1997.

158. P. Price and M. Ostendorf. Combining linguistic with statistical methods in modeling prosody. In J.L. Morgan and K. Demuth, editors, *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1995.

159. K. Shirai and S. Furui. Editors of special issue on spoken dialogue. *Speech Communication*, 15(3-4), 1994.

160. ESCA. Proceedings of the ESCA Workshop on Prosody. Technical Report Working Papers 41, Lund University, Department of Linguistics, 1993.

161. R.J. Lickley. *Detecting Disfluency in Spontaneous Speech*. PhD thesis, University of Edinburgh, Scotland, 1994.

162. E.E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, U. Cal. Berkeley, 1994.

163. G.R. Doddington. Speaker recognition—identifying people by their voices. *Proc. IEEE*, 73(11):1651-1664, 1985.

164. S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 5(2):183-197, 1986.

165. S. Furui. An overview of speaker recognition technology. *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 1-9, 1994.

166. A.E. Rosenberg and F.K. Soong. Recent research in automatic speaker recognition. In S. Furui and M.M. Sondhi, editors, *Advances in Speech Processing*, pages 701-737. Marcel Dekker, New York, 1991.

167. J.M. Naik, L.P. Netsch, and G.R. Doddington. Speaker verification over long distance telephone lines. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, S10b.3:524-527, 1989.

168. F.K. Soong, A.E. Rosenberg, and B.H. Juang. A vector quantization approach to speaker recognition. *AT&T Technical Journal*, 66:14-26, 1987.

169. S. Furui. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York, 1989.

170. J.L. Flanagan. Technologies for multimedia communications. *Proceedings of the IEEE*, 82(4):590-603, 1994.

171. J. Mariani. Speech in Multimodal Communication. In J. Flanagan and H. Fujisaki, editors, *Speech Communication for the Next Decade: new directions of research, technological development and evolving applications*. Honolulu, December 2-6 1996.

172. M. Denis and M. Carfantan, editors. *Images et Langages: multimodalites et modelisation cognitive*. Proceedings Colloque CNRS Images et Langages, Paris, April 1-2 1993.

173. D.G. Stork and M.E. Hennecke. Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, October 14-16, 1996.

174. W.H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. JASA, 26:212-215, 1954.

175. Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36:314-331, 1979.

176. M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In N.M. Thalman and D. Thalman, editors, *Models and techniques in computer animation*, pages 139-156. Springer-Verlag, 1993.

177. A.J. Goldshen, O. Garcia, and E. Petajan. Continuous optical speech recognition. In *Proceedings of the 28th IEEE Asilomar Conference on Signals, Systems and Computers*, 1994.

178. E. Petajan, B. Bischoff, D. Bodoff, and N.M. Brooke. An improved automatic lipreading system to enhance speech recognition. *CHI'88*, pages 19-25, 1988.

179. M.T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, et al. Multimodal learning interfaces. In *Proceedings ARPA 1995 Spoken Language Systems Technology Workshop*, Austin, January 22-25 1995. Morgan Kaufmann Publishers.

180. R.R. Sarrukai and C. Hunter. Integration of eye fixation information with speech recognition systems. In *Eurospeech'97*, Rhodos, September 22-25 1997.

181. D.A. Berkley and J. Flanagan. HuMaNet: An experimental human/machine communication network based on ISDN. *AT&T Technical Journal*, 69:87-98, 1990.

182. J.L. Flanagan. Overview on Multimodality. In E..Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, editors, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, September 1997.

183. O. Stock. A third modality of natural language? In P. McKevitt, editor, *Integration of natural language and vision processing. Intelligent Multimedia*, volume II. Kluwer Academics Publishers, 1995.

184. J.L. Gauvain, J.J. Gangolf, and L. Lamel. Speech recognition for an Information Kiosk. In *ICSLP'96*, Philadelphia, 3-6 October, 1996.

185. H. Waclar, T. Kanade, M. Smith, and S. Stevens. Intelligent Access to Digital Video: the Informedia Project. *IEEE Computer*, 29(5), May 1996.