Applications of Voice Processing to Telecommunications

LAWRENCE R. RABINER, FELLOW, IEEE

The ways in which people communicate are changing rapidly. No longer is the conventional voice call over a wired network the only reasonable and reliable method of transmitting information. Instead, the options are many and diverse, ranging from voice calls over wireless networks, to video calls over the conventional wired network, ISDN video, FAX, e-mail, voice mail, beeper services, data services, audio teleconferencing, video teleconferencing, and so-called scribble phone service (transmission of arbitrary handwritten input). This revolution in communications is being fueled by several sources, including the availability of low-cost, low-power, computation in both DSP and RISC chips, larger and cheaper memory chips, improved algorithms for communications (e.g., modems, signaling) and signal processing, and finally the creation of world-wide standards for transmission, signal compression, and communication protocols. The broad goal of the communications revolution is to provide seamless and high-quality communications between people (or groups of people), anywhere, anytime, and at a reasonable price. Although there are many technologies that form the bases for the communications environment of the twenty-first century, one of the key technologies for making the vision a reality is voice processing. In this paper we attempt to show, by example, how voice processing has been applied to specific problems in telecommunications, and how it will grow to become an even more essential component of the communications systems of the twenty-first century.

I. Introduction

Although voice processing has been a topic of research for several decades, it is primarily in the past few years that the technology has matured sufficiently to leave the research laboratory and enter the real world for a wide range of applications. There are several reasons why this has occurred. These include the rapid growth in computing capability provided by modern digital signal processing (DSP) chips, the sharp decrease in cost of computation and memory, and the increased emphasis on providing high-quality communications services.

The broad goal of voice processing technology is to help provide seamless, high-quality voice communications between people (or groups of people), anywhere, anytime, and at a reasonable price. To attain this goal requires significant advances in several technologies and in several specific application areas. Among these areas are the following:

1) **wireless communications**—both indoor and outdoor. This will provide the underpinning for *anywhere, anytime* communications capability.
2) **audio/video teleconferencing**—this will enable groups of people to communicate with each other in a manner so that the *groups* feel they are in a common environment (e.g., the same room).
3) **eyes-free, hands-free communication and control**—enabling the user to freely communicate without having to have physical contact with the communication device either for controlling the communication flow (i.e., initiating the call, transferring it to another number, etc.), or for communicating with the other party. This provides a *seamless* environment where control and communication are handled identically—i.e., by voice processing methods.
4) **acoustic echo cancellation**—providing an echo-free communications environment so that both parties perceive *high-quality* voice (and ultimately video) communications.
5) **smart microphones/loudspeakers**—providing a means of tracking individual talkers, adapting to any acoustic environment, and giving optimum signal-to-noise ratios while retaining spatial information about the sound field. Such smart acoustic transducers provide a *seamless* method of handling variability in both sound sources (e.g., multiple talkers in a group conversation), and in noise backgrounds.
6) **algorithmic advances and DSP implementation**—this provides a means for implementing the required voice processing technology on *reasonable cost* and *reasonable power* platforms while maintaining the high performance required for telecommunications applications.

Taken together, the vision of a Personal Information Terminal (PIT), in which a broad range of communications, computational, and signal processing capability is provided in a portable, wireless, device, becomes a reality. Even today the beginnings of such devices exist in the EO 440 Personal Communicator, and the Apple Newton Personal

Manuscript received August 10, 1993; revised October 28, 1993.
The author is with AT&T Bell Laboratories, Information Principles Research, Murray Hill, NJ 07974.
IEEE Log Number 9214857.

0018–9219/94$04.00 © 1994 IEEE

Terminal. Over time, the sophistication and capability of these devices will increase, with the devices becoming smaller in size, lower in weight, with longer life per battery charge, while providing integrated voice, video, e-mail, FAX, beeper, and data service over wireless channels. At the present time, however, although our voice processing capability falls far short of our vision, it is still extremely impressive when one considers the range of applications, within the telecommunications area, which have evolved. It is the purpose of this paper to discuss the individual areas of voice processing technology, to review the current status of the technology, and to show the extent to which applications have successfully been developed.

II. Areas of Voice Processing

The broad area of voice processing can be broken down into several individual areas, according to both applications and technology. These include:

1) **speech coding**—the process of compressing the information in a speech signal so as to either transmit it (or store it) economically over a channel whose bandwidth is significantly smaller than that of the uncompressed signal.
2) **speech synthesis**—the process of creating a synthetic replica of a speech signal so as to transmit a message from a machine to a person, with the purpose of conveying the information in the message.
3) **speech recognition**—the process of extracting the message information in a speech signal so as to control the actions of a machine in response to spoken commands.
4) **speaker recognition**—the process of either identifying or verifying a speaker for the purpose of restricting access to information (e.g., personal or private records), networks (computer, PBX), or physical premises.
5) **spoken language translation**—the process of recognizing the speech of a person talking in one language, translating the message content to a second language, and synthesizing an appropriate message in the second language, for the purpose of providing two-way communication between people who do not speak the same language.
6) **spoken language identification**—the process of identifying the language a person is speaking in, from the speech of that person.

In addition to each of these technology areas, work in the areas of speech analysis, hearing, and electroacoustics often forms the basis for the methods that are used to implement individual applications. Rather than discuss these areas of research individually, we will generally refer to them as they apply in individual situations.

In the remainder of this paper we discuss each area of voice processing technology separately. We first review the broad goals of each area, followed by a discussion of the basic processing techniques, ending with an evaluation of the current capability of the technology. We then discuss both typical and specific applications of the technology to telecommunications with the goal of focusing on the strengths and limitations of each technology.

III. Speech Coding [1]–[7]

Speech coding technology is used for both efficient transmission and storage of speech. For transmission applications the goal is to conserve bandwidth or bit rate, while maintaining adequate voice quality. For storage applications the goal is to maintain a desired level of voice quality at the lowest possible bit rate.

Speech coding plays a major role in three broad areas; namely, the wired telephone network, the wireless network (including cordless and cellular), and for voice security for both privacy (low level of security) and encryption (high level of security). Within the wired network the requirements on speech coding are rather tight with strong restrictions on quality, delay, and complexity. Within the wireless network, because of the noisy environments that are often encountered, the requirements on quality and delay are often relaxed; however, because of limited channel capacity the requirements on bit rate are generally tighter (i.e., lower bit rate is required) than for the wired network. Finally, for security applications, the requirements on quality, delay, and complexity are generally quite lax. This is because secure speech coding is often a requirement on low-bandwidth channels (e.g., military communications) where the available bit rate is relatively low. Hence, lower quality, long delay, low bit rate algorithms have generally been used for these applications.

Although we will discuss specific applications of speech coding later in this section, it is worth mentioning several broad classes of applications (beyond those used in transmission) of speech coding in the area of storage and teleconferencing. One of the largest application areas for speech coding is in voice messaging and voice mail, whereby a voice message is sent to a voice mailbox (either individually owned, or as part of a network), stored in coded form, and delivered to the intended recipient(s) when he or she is ready to receive it. Another important and growing application is the area of voice response systems whereby, usually in response to touch-tone input from a telephone (or ultimately in response to voice input commands), the system speaks out a coded message and thereby maintains a dialog with the user. Such voice response systems are being used as front-end processors for telephone queries to most major corporations and businesses. The field of digital coding of wideband speech signals is an emerging area of coding—especially for applications such as digital audio broadcasting (DAB) of compact disk (CD) audio over frequency modulation (FM) channels, and for surround sound for high-definition television (HDTV). Another interesting application is digital telephone answering machines where the usual tape recorder for storing of messages has been replaced by solid-state memory, thereby eliminating the problems associated with mechanical parts. Finally, the area of teleconferencing of coded wideband speech (in

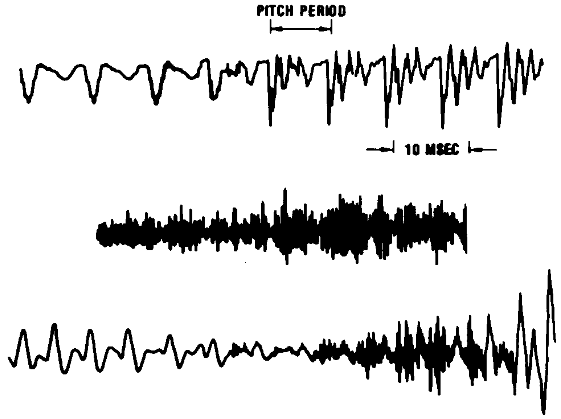


Fig. 1. Typical speech waveforms: upper panel: voiced speech; middle panel: unvoiced speech; lower panel: section of speech with both voiced and unvoiced sounds.

concert with coded video) is rapidly becoming used within businesses and should appear in the home and in wireless systems within the next few years.

A. Basic Principles of Speech Coding [3]

Figure 1 shows typical examples of speech waveforms. The top trace in Fig. 1 shows a voiced speech section (e.g., a vowel-like sound) waveform produced by modulating puffs of air (created by the vibrating vocal cords) by the vocal tract shape corresponding to the sounds being spoken. For such voiced waveforms, we see a quasiperiodicity of the signal (over periods of 10 ms) as well as a slowly changing waveform character. The middle trace in Fig. 1 shows the waveform of an unvoiced speech section (e.g., a sibilant sound like "s" or "sh") that has a noiselike character with no periodicity (the vocal cords are not vibrating) and no slowly changing temporal characteristics. Finally, the bottom trace in Fig. 1 shows a section of a speech utterance waveform that consists of both voiced and unvoiced sounds.

The fundamental process of speech coding is conversion of a speech signal to a digital representation (a sequence of binary digits), as shown in Fig. 2. The simplest way of obtaining such a digital representation is by applying the sampling theorem directly. This means that we must sample the speech waveform at a rate of twice the highest frequency present in the signal, and then digitize the resulting samples to some desired degree of accuracy. For telephone bandwidth speech signals (4–kHz bandwidth nominally), we need a sampling rate of at least 8000 samples/s and an encoding rate of 16 b/sample to maintain very-high signal-to-noise ratio in the digital representation. Hence, a total of 128 000 b/s is required for a high-quality digital representation of telephone bandwidth speech.

One of the goals of speech coding is to provide high quality at bit rates significantly below those implied by the sampling theorem. To achieve this goal we must exploit one or more of the special properties of speech signals to reduce the bit rate. Figure 3 illustrates a couple of the properties of speech waveforms that can be exploited; namely, the use of adaptive quantizers, whose characteristics vary over time, to match the dynamic range variations of the speech signal (contrast regions S and A in Fig. 3), and the use of time-varying filters to exploit both the short-time (within a single period) and long-time (across multiple periods) correlations of the signal. Figure 4 illustrates yet another important property of speech coding based on the fact that the decoded speech waveform is being perceived by a human listener [6]. The coding method can take advantage of well-understood properties of human hearing; namely, that noise (in this case quantization noise of the coding procedure) can be masked (perceptually hidden) by the speech signal if the spectral level of the noise is below the spectral level of the speech. As shown in the left side of Fig. 4, quantization noise is typically a flat-spectrum signal (i.e., random and uncorrelated with the speech) whose level often exceeds that of the speech and is therefore perceptually audible. By appropriately shaping the noise so that its spectrum matches that of the speech, as in the right side of Fig. 4, its level can be made to fall below the speech level at all frequencies of interest, thereby making it perceptually inaudible.

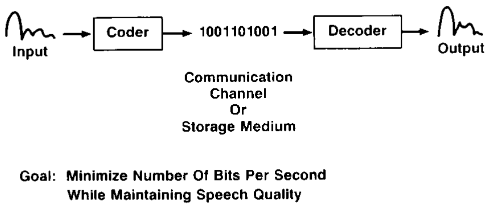

Input
Coder
1001101001
Decoder
Output
Communication
Channel
Or
Storage Medium
Goal: Minimize Number Of Bits Per Second
While Maintaining Speech Quality

Fig. 2. Conversion of analog speech signal to a digital representation via a coder and back to analog via a decoder.

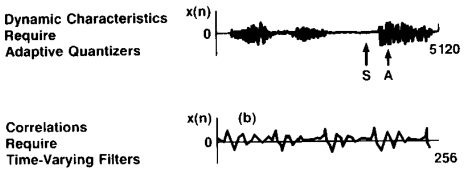


Fig. 3. Illustration of the need for an adaptive quantizer to efficiently code the highly variable dynamic range of speech, and a time-varying filter to code the local speech correlation.

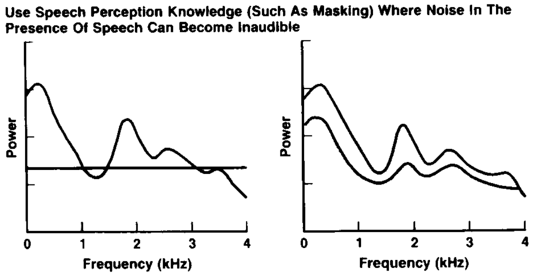


Fig. 4. Illustration of how the use of perceptual criteria (masking) can convert an audible white noise signal into an inaudible shaped noise signal with the same noise power.

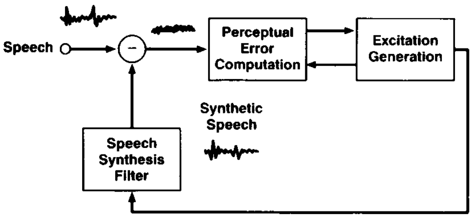


Fig. 5 Block diagram of a perceptually driven speech analysis/synthesis system.

Based on the above discussion, it can be seen that a typical speech coder consists of two modules; namely, an analysis module (called an analyzer) which extracts, from the speech waveform, the time-varying excitation waveform and the time-varying filter parameters, and a synthesis module (called a synthesizer) which recreates the "best" (in a perceptual sense) match to the original speech waveform. Fig. 5 shows a block diagram of such an analysis–synthesis approach to coding. The difference between the original speech signal and the output of the speech synthesis filter (the so-called error signal or the quantization noise of the coder) is perceptually weighted and minimized by adjusting parameters of the synthesis model, e.g., the excitation and the time-varying filters. Figure 6 shows several synthesis models which have been applied to speech coding, including the LPC vocoder model [1], the multipulse model, and the stochastic model [2], [5]. The vocoder (voice coder) model, the traditional speech synthesis model, categorically classifies the excitation signal as either voiced speech (with a quasiperiodic pulse train excitation signal) or unvoiced speech (with a random noise excitation signal), and feeds this switched excitation into an LPC (linear predictive coding) all-pole filter which models the time-varying spectral envelope of the speech signal. The multipulse model treats the excitation strictly as a sequence of pulses (typically 4 pulses every 5 ms) whose positions and amplitudes are determined automatically from the speech signal. The long delay correlation filter converts the excitation pulse train into a quasiperiodic signal, for voiced speech, and into a noise-like signal, for unvoiced speech, and the short-delay correlation filter models the spectral envelope in much the same manner as the LPC all-pole filter of the vocoder model. Finally, the stochastic model represents the excitation as the sequence, drawn from a stochastic codebook of random sequences, which, in conjunction with the long-delay correlation filter and the short-delay correlation filter, best matches the original speech signal (in a perceptually weighted sense).

Block diagrams of the full multipulse and stochastic coders[1] (called MPLPC and CELP for multipulse linear predictive coder and code excited linear prediction, respectively) are given in Figs. 7 and 8. For the multipulse

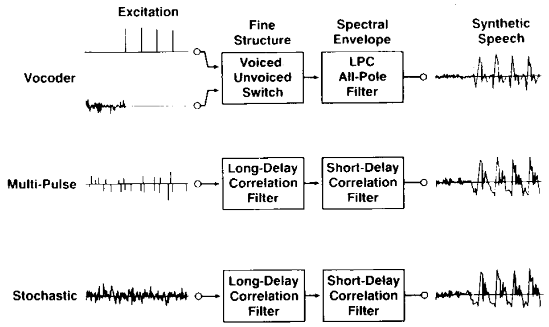


Fig. 6. Hierarchy of speech synthesis models, including the classic vocodor model, the multipulse model, and the stochastic (code-excited) model.

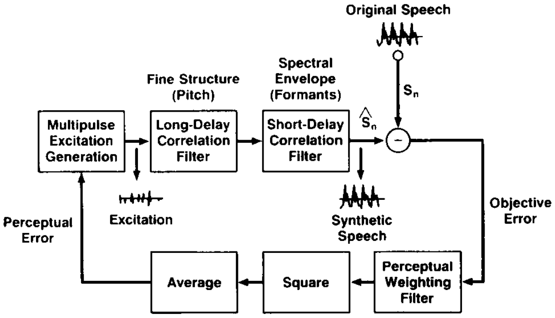

Original Speech
Fine Structure (Pitch)
Spectral Envelope (Formants)
S_n
$\hat{S}_n$
Multipulse Excitation Generation
Long-Delay Correlation Filter
Short-Delay Correlation Filter
Excitation
Synthetic Speech
Perceptual Error
Objective Error
Average
Square
Perceptual Weighting Filter

Fig. 7. Block diagram of the complete multipulse speech analysis/synthesis system.

coder, the difference between the original speech s_n and the current synthetic signal $\hat{s}_n$ (called the objective error in this figure) is perceptually weighted, squared, and averaged over a few 10's of milliseconds to give the perceptual error, which is then used to derive an improved excitation signal set of pulses for the synthesizer (i.e., the upper blocks in Fig. 7). This process is iterated, for each 5-ms frame, until the difference in the perceptual error from iteration to iteration becomes negligible. Similarly, for the stochastic coder of Fig. 8, each codebook excitation sequence is used as input to the synthesizer. For each 5-ms frame, the coder chooses the codebook sequence which gives the minimum perceptual error over all codebook entries.

By way of example, Fig. 9 shows waveforms, from the stochastic synthesis model, for the first 100 ms of a typical speech waveform. Shown in the figure are the stochastic excitation sequence (the top panel), the output of the long-term correlation filter (the second panel), the decoded speech (the output of the short-term correlation filter), and the original speech signal for comparison purposes. The high-quality reproduction capability of the stochastic model is clearly seen in this figure.

B. Bandwidth of Speech and Audio Signals

A key factor in determining the number of bits per second required to code a speech (or audio) signal is the

[1] Although originally the stochastic coders used random codebooks, most practical systems use codebooks derived from a training set of excitation vectors. Hence, strictly speaking, such coders are not stochastic. For historical reasons we will continue to refer to them as stochastic coders.

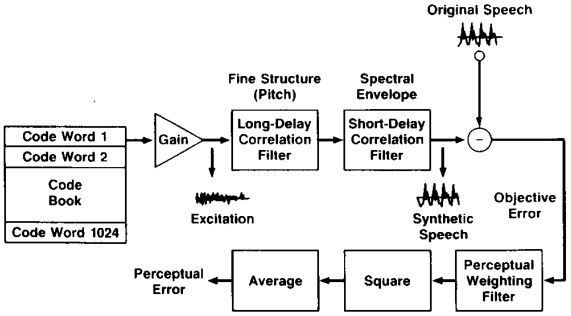


Fig. 8. Block diagram of the complete stochastic speech analysis/synthesis system.

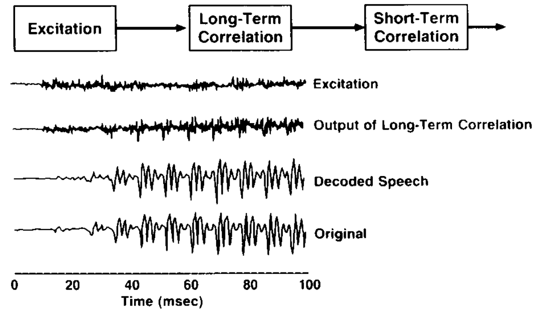


Fig. 9. Typical waveforms of various input and output nodes of the stochastic coder.

signal bandwidth. Figure 10 shows a plot of speech and audio signal bandwidth for four conventional transmission and/or broadcast modes; namely, conventional telephony, AM-radio, FM-radio, and compact disc (CD). Conventional telephone channels occupy a bandwidth from 200 to 3400 Hz; AM-radio extends the bandwidth on both ends of the spectrum to cover the band from 50 to 7000 Hz (this is also the bandwidth that most audio/video teleconferencing systems use for transmission of wideband speech); FM-radio extends the spectrum further (primarily for music) to the range 20 to 15 000 Hz; and the range for CD audio is from 10 to 20 000 Hz. Associated with these different transmission and broadcast modes are digital coding standards which we will discuss later in this section.

C. Evaluation of Speech and Audio Coders

All (digital) speech coders can be characterized in terms of four attributes; namely, bit rate, quality, signal delay, and complexity. The *bit rate* is a measure of how much the "speech model" has been exploited in the coder; the lower the bit rate, the greater the reliance on the speech production model. *Quality* is a measure of degradation of the coded speech signal and can be measured in terms of speech intelligibility and perceived speech naturalness. *Signal delay* is a measure of the duration of the speech signal used to estimate coder parameters reliably for both the encoder and the decoder, plus any delay inherent in the transmission channel. (Overall coder delay is the sum of the encoder delay, the decoder delay, and the delay in the transmission channel.) Generally the longer the allowed delay in the coder, the better the coder can estimate the synthesis parameters. However, long delays (on the order of 100 ms) are often perceived as quality impairments and sometimes even as echo in a two-way communications systems with feedback. Finally, *complexity* is a measure of computation required to implement the coder in digital signal processing (DSP) hardware.

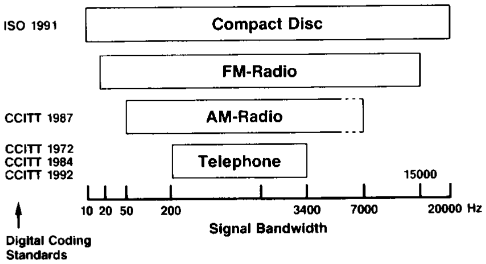


Fig. 10. Plot of speech and audio frequency bands for telephone, AM-radio, FM-radio, and compact disc audio signals.

The "ideal" speech coder has a low bit rate, high perceived quality, low signal delay, and low complexity. No ideal coder as yet exists with all these attributes. Real coders make tradeoffs among these attributes, e.g., trading off higher quality for increased bit rate, increased delay, or increased complexity.

To illustrate the current status of quality of telephone bandwidth coders, Figs. 11 and 12 show plots of speech intelligibility (as measured in terms of diagnostic rhyme test (DRT) scores), and speech quality (as measured in terms of mean opinion scores (MOS)) for a range of coders spanning bit rates from 64 kb/s down to 2.4 kb/s. (Also included in these figures are scores for uncoded telephone bandwidth natural speech.) The coders used in these tests included:

1) μ-law pulse code modulation (PCM) at 64 kb/s
2) adaptive differential pulse code modulation (ADPCM) at 32 kb/s
3) low delay code-excited linear prediction (LD-CELP) at 16 kb/s
4) vector sum excitation linear prediction (VSELP) at 8 kb/s (more precisely 7.950 kb/s)
5) code excited linear prediction (CELP) at 4.8 kb/s
6) linear predictive coding (LPC10 E) at 2.4 kb/s

The PCM and ADPCM coders are simple waveform coders with fixed or adaptive quantizers; the LD-CELP, VSELP, and CELP coders are stochastic coders; the LPC10 E coder is a US Government standard version of a vocoder model.

The DRT test measures intelligibility of speech in terms of distinguishing minimally distinct pairs of rhyming words, e.g., /bat/ and /pat/. It can be seen from Fig. 11 that the intelligibility scores for coders with bit rates of from 64 down to 4.8 kb/s are essentially identical, and only slightly lower than that of natural speech. At 2.4 kb/s a further slight degradation in intelligibility is observed. However, for the most part, all the coders maintain high DRT scores.

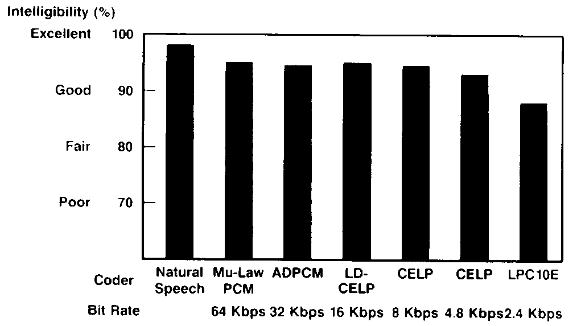


Fig. 11. Speech intelligibility scores in terms of diagnostic rhyme test (DRT) of several coders as a function of bit rate (PCM: pulse-code modulation; ADPCM: adaptive differential PCM; CELP: code-excited linear prediction; LD-CELP: low-delay CELP; LPC10 E: linear prediction coding).

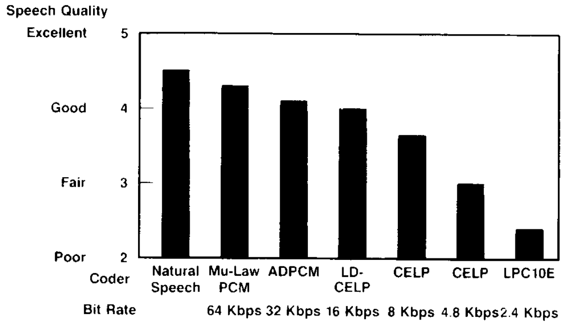


Fig. 12. Speech quality mean opinion scores of several coders as a function of bit rate.

The MOS test of speech quality uses a 5-point rating scale, with the attributes:

1) 5, excellent quality, no noticeable impairments
2) 4, good quality, only very slight impairments
3) 3, fair quality, noticeable but acceptable impairments
4) 2, poor quality, strong impairments
5) 1, bad quality, highly degraded speech

MOS scores are derived by averaging the responses of a large number of listeners, and are highly variable from test to test. To reduce the variability, MOS tests generally use a high-quality speech signal (either original speech or high-quality coded speech) as an anchor to stabilize the judgements of quality of the coded speech signals. Typical MOS scores for the high-quality anchor signals range from 4.0 to 4.5.

As can be seen in Fig. 12, there are significant differences in MOS scores among the different telephone bandwidth coders. The MOS score for the natural speech is 4.5, and all coders with bit rates from 16 to 64 kb/s achieved MOS scores of 4.0 or higher. Such high MOS scores are considered both necessary and sufficient for network applications of coders (e.g., transmission of speech) in which very high quality is required. At 8 kb/s, the MOS score of VSELP falls to 3.8, slightly below the level required for network applications, but quite useful in the noisier cellular network. At 4.8 and 2.4 kb/s, the MOS scores of the coders fall in the range of 2.0–3.0; such coders are acceptable primarily for military applications in which low bit rate is essential for secure (encrypted) communications.

D. Applications of Telephone Bandwidth Speech Coders in Telecommunications

There are four broad areas of applications of telephone bandwidth speech coders (outside of direct network transmission of coded speech) in telecommunications, namely:

1) voice messaging, including voice mail systems of all types
2) voice response, including coded messages in response to user queries via touch-tone or speech (recognition), and various information retrieval services. Voice response includes applications that answer as well as originate calls, and may use audiotex.
3) digital telephone answering machines, including coded prompts for time-of-day/date stamping of incoming messages, and coding of incoming messages
4) security devices, for encryption of sensitive voice information and transmission over channels of limited bandwidth.

In the following sections we discuss each of these areas in more detail.

1) Voice Messaging: Voice messaging is the technology to create, store, transmit, and deliver messages in voice form to either a personal voice mail box, or a network mail box for delivery at a later time. The fundamental premise behind voice messaging is that the majority of voice calls are fundamentally one-way information flow calls, and therefore do not need a network connection between two or more parties with the ensuing dialogue.

The advantages of voice messaging (over standard dialed-up calls) are as follows:

1) Increased messaging efficiency and accuracy; since voice messages are one-way calls, the lengths of these calls are significantly less than standard two-way calls—hence the amount of time it takes to "get across the message" is less than for standard telecommunications. This often leads to reduced phone bills for the caller and reduced load on the telecommunications network.
2) You reach the party you want, not a secretary; this feature gives everyone equal access to busy executives, business colleagues, etc., and provides the opportunity to present ideas that are unfiltered by an executive assistant or other intermediary.
3) Elimination of telephone call interruptions; all incoming messages can be queued up and answered in one session, rather than intermittently during the course of a day.
4) Ability to send messages without regard to time; this feature is essential to business travelers, colleagues in different time zones, etc., since a voice message can

be sent when the person is ready to send it, without having to wait for when the recipient is available to receive it.

5) Ability to forward messages; this feature enables the recipient of the voice message to decide the right person (or persons) to whom the voice message should go to.
6) Ability to easily retrieve messages anywhere, anytime; this feature provides the convenience of communications whenever the recipient of the voice message is ready to act on the message, and wherever the person is located at the time.
7) Ability to broadcast messages; this feature allows the recipient of the message to share its contents with a broader audience without interpretation or modification of the original message content.
8) Privacy; this feature provides the security that the voice message can only be retrieved by the person to whom the voice message was sent.
9) Ability to reach party in a single call attempt; this eliminates the frustration of waiting for a busy line to become free, or of being routed from one party to another in order to try to get the attention of an executive of some company, or of telephone tag.
10) Eliminates queues in calling; this feature is another time-saving feature since the call is handled by the voice messaging center which has far more lines (hence less waiting) than most ordinary businesses provide.

Overall, these very significant advantages of voice messaging have enabled voice mail to replace the written memo in many instances of daily communications in the workplace.

Types of voice messaging systems: Voice messaging systems all fundamentally perform the same function—namely, preserving a voice message in digitally coded format for non-real-time access at the convenience of the party to whom the voice message is addressed. Voice messaging systems differ in several aspects, including the physical location of the messaging system, the type of coding used to store the messages, the location of the voice mailboxes, and the degree of networking that the system is capable of addressing.

A simple classification scheme for voice messaging systems is in terms of where the processing for coding, storing, transmission, and decoding of messages takes place. There are four broad classes of systems, based on these criteria, including:

1) **stand-alone systems.** This type of system is independent of the network used for transmission of the messages and usually consists of hardware and software for coding, storing, and decoding messages located on customer premises. Such systems generally provide a range of voice messaging services such as the ability to broadcast voice messages to specified groups of people at specified times, dates, etc. Typical vendors of these systems include Octel, VMX, and Centigram.
2) **PBX-based systems.** This type of system is attached to and slave of a PBX. Hence the system is physically located on the premises of the network provider of the transmission and switching equipment. Full functionality of voice messaging services is usually available with these systems including essentially unlimited mailbox capability, mailbox searching for individual stored messages, message cataloging and archiving, variable rate playback of messages, multiple mailbox per user capability, etc. By the end of 1992, it was shown that about 50% of new PBX sales included voice-mail capability. Typical providers of PBX-based voice messaging systems include AT&T (Audix), Northern Telecom, and ROLM.
3) **PC-based systems.** This type of system is geared to an individual user and provides the key feature of integration of electronic mail (e-mail) with voice mail, so that all storage, coding, and decoding of the voice mail is performed on the local PC. Hence the cost of these individual voice messaging systems is low. A typical PC provider of voice mail is NOVELL.
4) **Service bureau messaging.** This last type of voice messaging system is a reselling of systems by third party organizations. Instead of having to purchase their own voice messaging system, companies essentially "rent" voice mail boxes and associated services from the service bureau. Hence individual users can select the services they need, and easily change capability over time as the demand for such services changes with growth (or slow down) of business. The major service bureau for messaging is TIGON with on the order of 100 000 users by the end of 1992; the RBOC's are also beginning to provide this service.

2) Voice Response Systems: Voice response systems consist primarily of prerecorded and digitally coded announcements, and words and phrases, which are used to provide voice responses to customers from queries made via telephone connections to either companies or specific customer-accessible databases. There are two broad classes of voice response system; namely, automated attendants and interactive voice response (IVR) systems.

Automated attendants provide either voice routing of calls (via either touch-tone or spoken queries), or voice routing of voice messages (again via either touch-tone or spoken queries). Hence the typical automated attendant, in response to a customer dialing into a corporation, provides a voice response prompt asking the customer to enter a code for the type of service (or for an individual) requested. Based on the entered code, either a live attendant is provided, or additional voice prompts are used to guide the customer.

Interactive voice response systems are used either to dispense specific repetitive information (e.g., weather in different cities, traffic conditions on highways, airplane arrival and departure times, etc.), or to provide user-requested information as retrieved from a dynamic database (e.g., stock price quotations, airline fares, availability of tickets to specific theater shows, etc.).

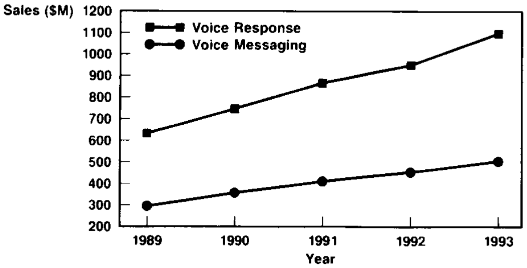


Fig. 13. Plot of the growth in sales of voice processing equipment from 1989 to 1993.

IVR systems have grown rapidly in the past few years and have been installed in virtually every call center across the United States. The reason for the growth in the use and popularity of IVR systems is that (as estimated by Travelers Insurance) about 60% of all calls to major corporations do not require simultaneous conversation (i.e., need to access a live attendant).

The benefits of interactive voice response systems include the following:

1) The desired information can be obtained more rapidly and significantly less expensively since the systems do not require live attendants to be available for most queries.
2) Services can be available 24 h a day; hence the user can avail himself (or herself) of the service whenever he (or she) desires.
3) The call can be kept private since live attendants are not required.
4) The information provided can be easily kept up-to-date and consistent across a wide range of services.
5) Allows live attendants to concentrate on calls requiring personal attention.
6) Provides the capability of multilingual systems so that the information and voice prompts can be provided in the language of choice of the user.

To illustrate the ubiquity of voice response applications, consider the following generic applications that currently use voice response to interact with customers:

1) **call screening.** For this application the system asks the customer to provide specific information (e.g., credit card number, type of interaction, etc.) so as to give the live attendant enough information to help the customer more rapidly.
2) **call scheduling.** For this application the system acts as a reminder service to help the customer wake up at specified times, remember appointments, birthdays, anniversaries, or even to schedule work tasks such as servicing or repairing equipment at specified times or as needed.
3) **call processing.** For this application the system provides either call forwarding (if the customers wants to speak to a particular individual), or routing to an alternative party when the callee is unavailable.

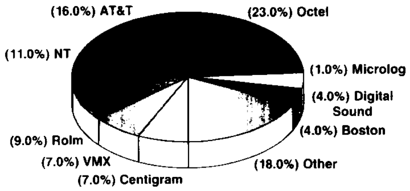


Fig. 14. Distribution of the share of the voice messaging market in 1991.

4) **order tracking.** For this application either the customer, or the salesperson, can determine the status of individual orders by interactively accessing service order records.
5) **business locator service.** For this application the system provides the address and telephone number of the nearest location of a specified business (e.g., car dealer, department store, Pizza Hut, etc.) to the customer.
6) **catalog services.** For this application the system provides access for ordering from a range of catalogs—either user-specified, or based on a profile of the caller's interests.

Some specific industries which use IVR systems as part of their everyday business include:

1) banking: for access to accounts, loan information;
2) education: for student registration, grade reports, account balances, tuition payments;
3) cable TV: for ordering of pay-per-view programs;
4) airlines: for flight status, routing and scheduling of flight plans, preflight check-in;
5) insurance: for routing of calls, updates on checks;
6) utilities: for service and billing information, handling of work orders;
7) transportation: for status of shipments;
8) medical: to verify training and employment history of job applicants;
9) retail: for automation of help desk services;
10) financial: to provide stock prices, information on trades, and for alerting services;
11) shipping: to provide status of shipments.

3) Markets for Voice Messaging and Voice Response: To illustrate the growth in the markets for voice messaging and voice response systems, Fig. 13 shows a graph of the sales of such systems in the US from 1989 through 1993 (estimated for 1993). It can be seen that the market for voice messaging systems has risen from about $290 M in 1989 to about $500 M in 1993 (estimated), whereas the voice response market sales has risen from about $630 M in 1989 to almost $1.1 B in 1993 (estimated).

A breakdown of the US market for Voice Messaging market for 1991 is shown in Fig. 14. The market leader that year was Octel with 23% of sales, followed by AT&T with 16% of sales, then Northern Telecom (11%), ROLM (9%), VMX (7%), Centigram (7%), and others providing the remaining 27% of the business. It is seen that no single company dominates the Voice Messaging market.

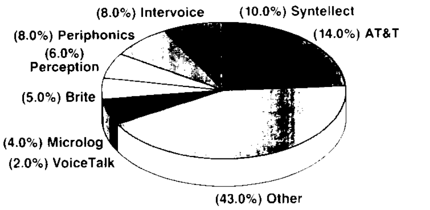


Fig. 15. Distribution of the share of the voice response market in 1991.

Similarly, Fig. 15 shows a breakdown of the US market for Voice Response market for 1991. Here the market leader was AT&T (Conversant Systems) with 14% of sales, followed by Syntellect (10%), Intervoice (8%), Periphonics (8%), Perception (6%), Brite (5%), with 49% of sales going to other providers. It is again seen that no single company dominates the Voice Response market.

E. Telephone Answering Machines

Another evolving class of applications of speech coders is in digital telephone answering machines. With the advent of large, inexpensive, solid-state memories (e.g., 4 and 16 Mb), and with appropriate low rate speech coders (e.g., 6.6- to 13.0-kb/s range) on the order of 5 min of coded speech can be stored on a single 4-Mb chip, and on the order of 20 min of coded speech can be stored on a 16-Mb chip. Hence the usual 30-min tape drive (with all the problems associated with mechanical drives, tape dropouts, tape capstans, etc.) can be effectively replaced by a speech coding/decoding chip (usually a low-cost DSP chip) and one or more memory chips to store both the voice prompts and the incoming messages.

By way of example, Fig. 16 shows a picture of the AT&T 1343 Digital Telephone Answering Machine. This machine has multiple voice prompts, and storage for both announcements (using LPC10 E coding at 2.4 kb/s) and 28 min of incoming messages [using RPE-LTP (Regular Pulse Excitation with Long Term Prediction) coding at 13.0 kb/s].

F. Telephone Security Devices

One last general area of applications of voice coding in telecommunications is the area of security. Such systems both encode the telephone speech digitally and encrypt the resulting bit stream using some data encryption standard such as DES (data encryption standard). Because of the requirements (as established by the usage of such devices by government and military agencies) that such security devices be capable of communicating over virtually any military channel, the maximum allowable speech data rates are in the 2.4–4.8-kb/s range. Two generations of these security devices have evolved, the first resulting in a dual-mode system capable of transmission at both 2.4 and 4.8 kb/s (the so-called Secure Telephone Unit (STU III) device), and the second (called the Secure Telephone Device 3600) running just at a 4.8-kb/s rate. The STU-III uses LPC10 E coding at 2.4 kb/s and CELP coding at 4.8 kb/s. The STD 3600 uses RCELP coding at 4.8 kb/s. Fig. 17 shows photographs of both the STU-III and the TSD-3600 devices, as produced by AT&T.

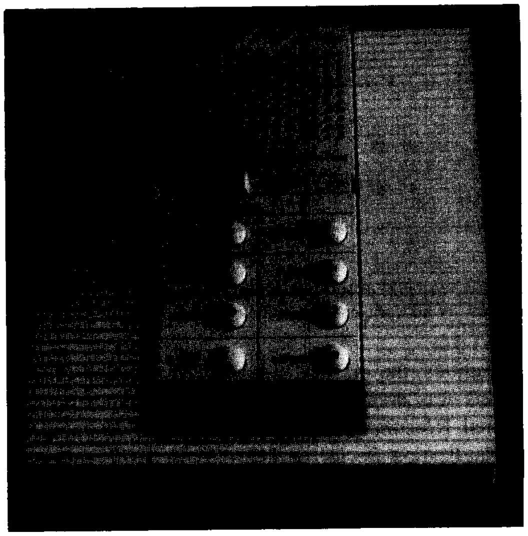


Fig. 16. Photograph of the AT&T 1343 digital Telephone Answering Machine.

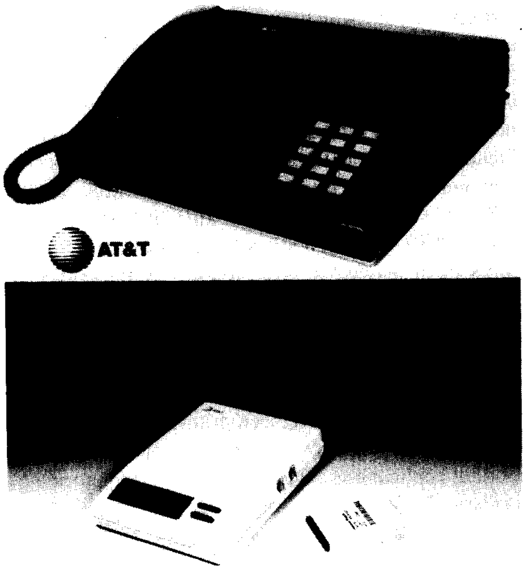


Fig. 17. Photographs of two telephone security devices, the AT&T STU-III terminal , and the AT&T TSD-3600.

G. Standards for Telephone Bandwidth Speech Coding

A key driving force in the widespread use of speech coding in telecommunications is the standardization of speech coding algorithms for interoperability in various transmission systems. Standards have been created for network applications, for mobile radio/cellular applications, and for secure voice applications. Figure 18 shows a plot which illustrates the standards which have been created in each of these areas. In the area of coding for network

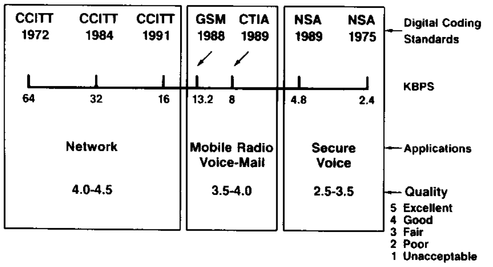


Fig. 18. The range of bit rates and MOS scores for voice coding standards for network, cellular, and security applications.

applications, the original μ-law PCM standard (G.711) at 64 kb/s was created by the CCITT (now renamed International Telecommunications Union—Telecommunications Standardization Sector or ITU-TS) in 1972, followed by the ADPCM standard (G.721) at 32 kb/s in 1984, and most recently, the LD-CELP standard (G.728) at 16 kb/s in 1991. In the cellular arena, the European digital standard (GSM) was created in 1988 with a 13.0-kb/s rate, and the VSELP Northern American digital standard (IS-54) at 8 kb/s was created in 1989. (There is also a Japanese standard (JDC), based on VSELP, which operates at 6.7 kb/s which is not shown in Fig. 18.) Currently, there is renewed activity for a new 8-kb/s standard as well as activity to create so-called half-rate digital standards in both Europe and North America but these standards have not yet been approved.

Finally, the National Security Agency (NSA), has created secure voice coding standards at both 2.4 kb/s (LPC10 E or FS-1015) in 1975, and at 4.8 kb/s (FS-1016) based on CELP in 1989. Although improved coding is used in newer security devices (e.g., RCELP, relaxed excitation CELP, in the TSD-3600), there are no standards, as yet, for these new algorithms.

H. Wideband Speech Coding

Until this point, we have been primarily discussing methods for coding telephone bandwidth speech. For many important applications a wider bandwidth is appropriate and necessary. These applications include:

1) Audio and video teleconferencing where broadened bandwidth (50–7000 Hz) provides improved sound quality, more presence of the speaker, and a more realistic rendering of the actual sound in a room.
2) Digital AM radio broadcasting where the 50–7000-Hz band is currently used for high-quality voice transmission.
3) High-fidelity telephony where broadcast-quality voice is transmitted over cables, fiber-optic networks, or even the local loop (after modification to eliminate the current bandlimiting networks).
4) Dual-language programming in audio and audio/video broadcasts of news, TV programs, closed-circuit lectures, etc.

Based on the growing needs of wideband speech in telecommunications, standard CELP methods have been applied and have been shown capable of providing high-quality speech (MOS scores of 4.0 or higher) in the 32–64-kb/s range. Current research is focusing on lowering the bit rate to 16 kb/s while maintaining high quality so as to provide audio/video teleconferencing at 128 kb/s with 112 kb/s provided for video coding, and 16 kb/s for high-quality audio coding.

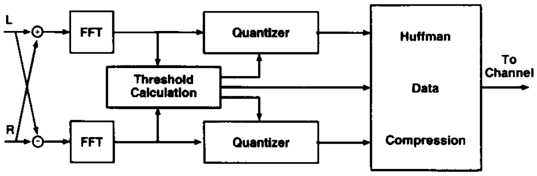


Fig. 19. Block diagram of PAC (Perceptual Audio Coder).

I. CD Audio Coding [4]

With the advent of mass marketed devices for digital coding and storage of high-fidelity audio, including the Compact Disc (CD), the digital audio tape (DAT), and most recently the minidisk (MD), and the digital compact cassette (DCC), the area of efficient digital coding of high-fidelity audio has become a topic of great interest and a great deal of activity. Also driving this activity is the need for a digital audio standard for the sound for high-definition TV (HDTV) and for digital audio broadcasting (DAB) of FM-channels.

To appreciate the importance of coding digital audio efficiently and with quality which is essentially indistinguishable from that of an original CD, consider the bit rate that current CD's use to code audio. The sampling rate of a CD is approximately 44.1 kHz and each sample (for both channels of a stereo broadcast) is coded with 16-b accuracy. Hence a total of 44.1 × 2 × 16 or 1.41 Mbps is used to code digital audio on a CD. Current state-of-the-art coding algorithms, such as the Perceptual Audio Coder or PAC developed at AT&T Bell Labs, are capable of coding 2 channels of digital audio at a total bit rate of 128 kb/s with essentially no loss in quality from that of the original CD coding [4].

A block diagram of the PAC coder is shown in Fig. 19. The stereo audio signal (left and right channels) is first spectrally analyzed, using a high-resolution FFT analysis of both the sum of the left and right channels, and the difference of these channels. Next, a perceptual threshold of noise (distortion) audibility is determined based on our understanding of masking versus frequency, resulting in a masking threshold curve of the type shown in Fig. 20, where the jagged curve is the spectral level of the signal and the staircase-like curve is the masking threshold below which noise (distortion) is inaudible. Next, the signals (both sum and difference) are quantized to the desired bit rate with the goal of shaping the quantization noise so that it falls below the masking threshold curve at all frequencies. Finally, the resulting digital bitstream is Huffman coded so as to compress it optimally, at the same time removing redundancy in the quantizer bit stream.

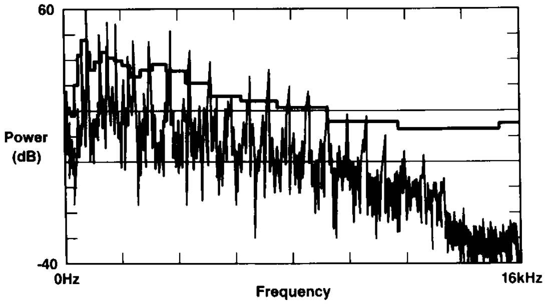


Fig. 20. Plot of the spectral energy density of a masked sound along with the perceptual masking threshold appropriate to this sound.

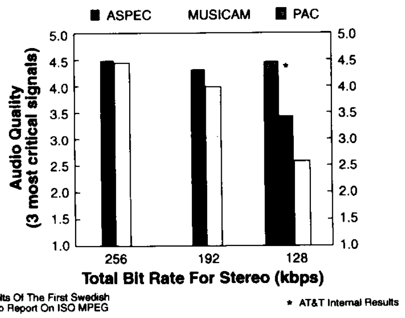


Fig. 21. Plot of MOS scores for several audio coders at 256-, 192-, and 128-kb/s total rates.

Although the exact details of the PAC coder differ somewhat from the description above, in essence PAC tries to code the audio signal so as to maximally exploit the perceptual masking of noise by strong audio signals in adjacent frequency bands. The success with which PAC (as well as an earlier AT&T coder called ASPEC, and a Philips coder called MUSICAM) can code digital audio at 128 kb/s is shown in Fig. 21, which shows MOS scores for these coders at total bit rates of 256, 192, and 128 kb/s. It can be seen that MOS scores of 4.0 or higher are obtained for both ASPEC and MUSICAM at the higher rates (256 and 192 kb/s); however, at 128 kb/s both these coders have MOS scores which are significantly below that of the original CD recording (typically around 4.5). Based on informal internal tests at AT&T, it can be seen that the PAC coder obtains an MOS of 4.5 at 128 kb/s. Since the MOS scores of ASPEC and MUSICAM were obtained in earlier tests, these coders may also have improved and obtained higher MOS scores at these rates.

J. Computational Requirements of Coders

A key requirement for most speech coders is that their computational requirements fall well within the range of modern digital signal processing (DSP) chips so that the coders can be implemented both inexpensively and efficiently. Table 1 shows a list of the computational requirements, as measured in direct implementations at AT&T Bell Laboratories (in terms of millions of instructions per second, or MIPS) for telephone bandwidth coders, wideband speech coders, and audio coders. Included in the list of coders are both standard coders (i.e., coders for which standards are in place) and more recent coders (such as the time–frequency interpolation coder (TFI) [7]) for which no standards exist.

For telephone bandwidth coders, it can be seen that as the compression ratio increases, the MIPS requirements generally also increase, sometimes disproportionately. Coders like the TFI coder often require far more computational capability than can be provided by even the most advanced DSP chip today. Hence a research goal is to reduce the MIPS requirements of algorithms while sacrificing as little quality of the coding as possible.

For both wideband speech, and audio coders, it can again be seen that the computational requirements of the algorithms are quite large, thereby severely limiting the utility of these coding algorithms to broad classes of applications. As capability of DSP chips increases, and as understanding of coding algorithms improves, the computational requirements of these advanced coders generally fall into the range for mass deployment.

IV. Speech Synthesis [8]–[17]

The goal of speech synthesis is to provide a broad range of capability for having a machine speak information (respond) to a user. A simple communications model of a system for voice access to information (e.g., databases) is shown in Fig. 22. It is assumed that a user wants to get information into or out of an existing database (e.g., pay bills, check account balances, etc.), and that the user has access to a touch-tone receiver (TTR) or an equivalent keyboard. (If this is not the case the user can enter requests for information using voice commands. We will discuss this alternative further in the next section of this paper.) The user enters requests for information, via the TTR Keypad, to a communications interface that transmits the request to a database manager. The requested information is sent back to the user (again through the communications interface) in the form of voice output (synthesized speech) as this is the only output modality supported on a standard TTR. Thus the key issue is how to convert the text equivalent of the database information efficiently to speech for different applications.

There are three factors affecting the way in which a synthesis system is implemented for different tasks; namely, the required quality of the synthetic speech, the range of speaking vocabulary, and the cost (complexity) of the synthesis software and hardware. Included in the cost factor is the storage costs for speech units (dyads, words, phrases), rules, and dictionaries, as well as the cost of the speech generation hardware.

A. Typical Applications of Speech Synthesis [11]

There are two broad classes of synthesis applications: those that require little or no user interaction and those that are highly user interactive. In the first class are a

Table 1 Computation Requirements for Speech and Audio Coders

Telephone Bandwidth Speech Coding (3.2-kHz Bandwidth)

Algorithm	Rate (kb/s)	Compression	MIPS	Application
PCM	128	1	0	network
Mu-Law	64	2	0	network
ADPCM	32	4	1	network
LD-CELP	16	8	50	network
LC-CELP	16	8	10	voice messaging coder
RPE-LTP	13.0	9.7	10	digital cellular
VSELP	8	16	24	digital cellular
CELP+	6.8	18.8	30	videophone/digital cellular
RCELP	4.8	26.7	16	Telephone Security Device
CELP	4.8	26.7	30	security
TFI	4	32	150	digital cellular
LPC10E	2.4	53.3	15	security
TFI	2.4	53.3	120	security

Wideband Speech Coding (7-kHz Bandwidth)

Algorithm	Rate (kb/s)	Compression	MIPS	Application
Uncompressed	256	1	0	teleconferencing
Subband Coder	64	4	10	teleconferencing
LD-CELP	32	8	100	teleconferencing
LD-CELP	16	16	400	teleconferencing

Audio Coding (20-kHz Bandwidth Two-Channel Stereo)

Algorithm	Rate (kb/s)	Compression	MIPS	Application
Uncompressed	1410	1	0	digital audio
ASPEC-Encoder	256	5.5	150	digital audio
-Decoder			25	digital audio
PAC-Encoder	128	11.0	230	digital audio
-Decoder			20	digital audio

broad range of telecommunications applications such as the Automatic Intercept System (AIS), in which a telephone call dialed to an inactive (or out-of-service) number is intercepted and an appropriate message with details about the problem is spoken back to the user; the class of "talking announcements" such as the local time or weather; and the class of entertainment and communications services such as dial-a-joke, dial-a-prayer, daily horoscope, etc. In almost all cases these noninteractive services are provided using the Interactive Voice Response (IVR) capability discussed in the previous section. In some cases, however, it is both convenient and more practical to use some type of text-to-speech (TTS) capability to provide the messages—especially when they change often in the course of a day.

Included in the second class of user interactive applications of speech synthesis are voice servers for reading e-mail and FAX messages over phone lines; standard database access services such as voice banking, stock price quotations, sports scores, flight information, etc.; and finally, services that require the ability to speak unlimited, unconstrained text as found in medical textbooks, legal volumes, and encyclopedias. In almost all cases these user interactive applications generally require full TTS capability.

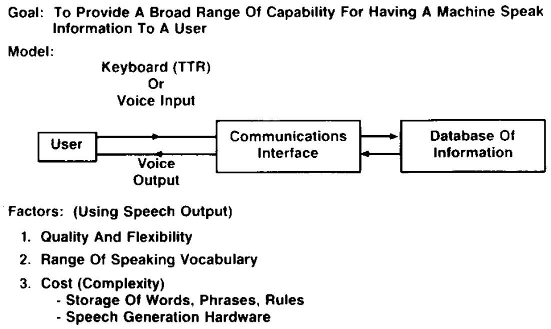

Goal: To Provide A Broad Range Of Capability For Having A Machine Speak Information To A User
Model:
Keyboard (TTR)
Or
Voice Input
User
Voice
Output
Communications Interface
Database Of Information
Factors: (Using Speech Output)
1. Quality And Flexibility
2. Range Of Speaking Vocabulary
3. Cost (Complexity)
- Storage Of Words, Phrases, Rules
- Speech Generation Hardware

Fig. 22. Model of communications system using speech synthesis to provide spoken messages to the user in response to TTR (touch-tone receiver) or voice input requests.

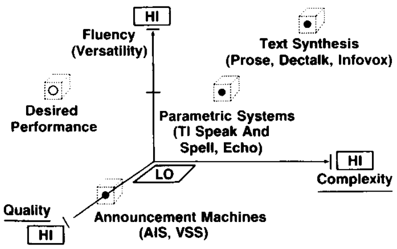


Fig. 23. Examples of several general classes of synthesis systems in terms of their quality, fluency, and complexity. (AIS: Automatic Intercept System; VSS: Voice Storage Services.)

B. Factors Affecting Speech Synthesis

There are three key factors which influence the use of speech synthesis systems for different applications; namely, the quality (measured in terms of intelligibility and naturalness) of the synthesized speech, the fluency of the spoken output (i.e., the ability to create messages with different vocabularies, emphasis, intonation, speed, etc.), and the complexity (as measured in terms of both storage and computation) of the synthesis hardware. To illustrate the interaction of these three factors for a range of systems, Fig. 23 shows a plot of where several current systems would fall in this three-dimensional space. The "ideal" synthesis system would provide high quality (the resulting speech is both highly intelligible and natural), high fluency (virtually any text message could be produced at the desired speaking rate and with the desired emphasis), and would be low cost (so that it would be cheap enough to integrate into any desired application). This ideal system is shown as the dotted point labeled "Desired Performance" in Fig. 23. Unfortunately, in the real world, there is no practical system with a performance that even comes close to the ideal. Instead, as shown in Fig. 23, there are three actual classes of systems: announcement machines (IVR systems), as might be used in the AIS application or for Voice Storage Services (VCS); parametric systems (as exemplified by the Speak-n-Spell toy introduced by Texas Instruments); and full TTS systems (such as those by Prose, DEC, Infovox, and AT&T Network Systems and Conversant Systems).

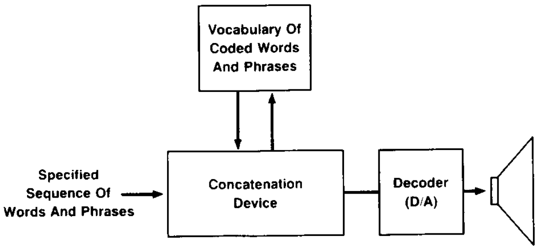

Vocabulary Of Coded Words And Phrases
Specified Sequence Of Words And Phrases
Concatenation Device
Decoder (D/A)

Fig. 24. Block diagram of simple voice response system (D/A: digital-to-analog).

The announcement machine systems provide high-quality synthesis (since they use prerecorded and digitized speech messages), but have low fluency (since they can only speak the prerecorded messages or trivial combinations of them), and low complexity. The parametric systems provide low-quality speech (highly synthetic sound), with medium fluency (they can speak a reasonable range of text), with low-to-medium complexity. Full TTS systems currently provide low to medium quality, with high fluency (they can speak any ASCII text), at relatively high complexity (primarily for units and dictionaries).

C. Speech Synthesis Technology

Speech synthesis systems can be realized as either simple concatenation systems, as shown in Fig. 24, or as full TTS systems, as shown in Fig. 25. The concatenation system has a stored vocabulary of prerecorded and digitally coded words and phrases. Based on user actions (e.g., dialing a disconnected telephone number), a request for a specified sequence of words and phrases is generated and sent to a concatenation device that retrieves from the digital store the coded versions of each of the required vocabulary items, concatenates the vocabulary items for the message, and sends the final result to a decoder that produces the analog speech heard by the user. Thus for the intercept message, "The number you have dialed, 555–1234, has been disconnected," the concatenation system retrieves, in sequence, the phrase "The number you have dialed," followed by each of the digits in the telephone number, followed by the phrase "has been disconnected." For naturalness, usually several versions of each digit are stored so that a digit at the beginning of the telephone number has a different duration and emphasis than the same digit in the middle or at the end of the telephone number.

For the full TTS system of Fig. 25, the desired message text is an arbitrary ASCII string (usually, but not always, with appropriate punctuation), so the first task of the system is to convert the text string to a sequence of phonetic symbols (indicative of the sounds to be spoken), along with a set of prosody markers (indicating the speed of the speech, the intonation, and the emphasis on certain words). This "text-to-sound/prosody" conversion involves a combination of linguistic analyses including dictionary lookup of word pronunciation and rules for exceptions and unusual cases, algorithms for generating appropriate word durations, and

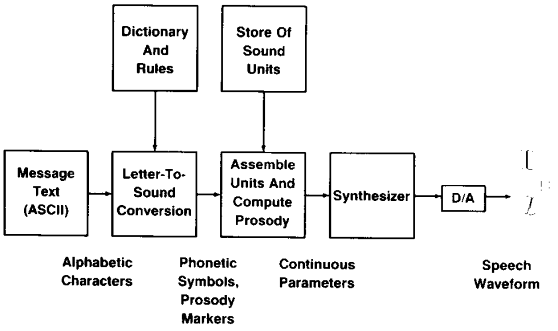


Fig. 25. Block diagram of full text-to-speech synthesis system.

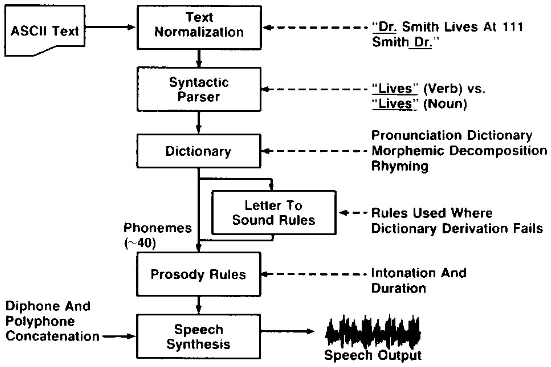


Fig. 26. Flow diagram of linguistic analyses of text used in the TTS system of Fig. 25.

algorithms for generating an appropriate pitch and loudness contour for the speech. Once the appropriate phonetic symbols and prosody markers have been determined, the next step in the TTS process is to assemble the appropriate speech units and compute the pitch and duration contours for the speech. To do this a store of elemental sound units is required. Creation of an appropriate set of these synthesis units is both time-consuming and difficult, as these units must be robust to different phonetic environments, yet must be rich enough to disambiguate sound combinations that are different in minimal ways. Experience with several AT&T TTS systems shows that sound inventories of from 2000 to 4000 dyad/polyad units (dyads are spectral representations of time slices from 2-phone sequences, polyads are spectral representations of time slices from sequences of 3 or more phones) are required for good-quality synthesis [15]. Other systems (e.g., Dectalk, Prose, Infovox) use formant synthesis methods, rather than storing dyad inventories [8], [9], [13], [14], [16]. The final steps in the TTS process are synthesis from spectral parameters appropriate to the sequence of synthesis units, and digital-to-analog (D/A) conversion of the resulting speech to render it useful for transmission back to the user.

D. TTS Text Analysis

In order to create the correct phonetic symbols for words from text, several text analysis procedures must be used so as to resolve ambiguity of several types. To illustrate these problems, consider the following sentences:

1) *Dr. Jonasz lives* on Bourban *St.* in *St.* Louis
2) The nine *lives* of Felix the cat of Segamore *Dr.*
3) The *prject* is *giong* well—at least until *BCSys* and *NCR* get their acts together.

Consider the ASCII text /Dr./ which is pronounced as /Doctor/ in Sentence 1 and as /drive/ in Sentence 2; or the text /St./ which is pronounced as /Street/ at the first occurrence in Sentence 1, and /Saint/ at the second occurrence. Other issues include words like /lives/ which are pronounced as "livz" when it is a verb in Sentence 1, and "laivz" when it is a noun in Sentence 2; acronyms like /BCSys/ and /NCR/ in Sentence 3 which are essentially undefined as to how they should be pronounced; typos like /prject/ and /giong/ in Sentence 3; foreign names like /Jonasz/; and finally unusual punctuation which often signals a pause or other durational cues.

In order to properly handle the text problems discussed above requires a series of linguistic analyses of the text, as outlined in Fig. 26. First the incoming ASCII text is normalized so as to expand common abbreviations like /Dr./ and /St./, and to expand punctuation, number sequences, dollar amounts, etc. The text normalization module also finds and marks acronyms for later analysis by the word pronouncing module. The normalized text is then fed into a syntactic parser which determines both the grammatical structure of the text and the part of speech (POS) for each word in order to resolve verb/noun or adjective/noun ambiguities, among others. A semantic analysis is also made to provide a representation of the meaning of the text to aid in word pronunciation. Next a series of phonetic, phonological, and morphological analyses of words is made to derive the appropriate set of word pronunciations based on both a large word dictionary and a set of letter-to-sound rules for cases where the dictionary derivation fails (e.g., proper nouns, foreign words). The current AT&T English word dictionary stores 57 000 different word pronunciations. Based on a prefix, root, and suffix analysis, the dictionary was compressed to 30 000 roots, which, when expanded, provided accurate coverage (in terms of word pronunciation) of 166 000 words. In addition to the root dictionary, a second dictionary of common names is used to aid in the pronunciation of proper names. This, along with appropriate letter-to-sound rules, provides good name pronunciation accuracy for a large percentage of common surnames in the United States.

To understand how difficult a problem accurate pronunciation of proper names can be, consider the following statistics from the RH Donnelly list of the 1.5 million most common surnames in the US. The 5000 most common names cover 59.1% of the 1.5 million list; the 50 000 most common names cover 83.2% of the list; the 200 000 most common names cover 93.0% of the list. Hence a small name dictionary, along with appropriate letter-to-sound rules, is a good compromise for accurate proper name pronunciation.

The final linguistic analysis that is performed in the TTS system is a discourse analysis to quantify relationships among sentences and ideas (in a paragraph). This type of analysis identifies places where pauses should be inserted, as well as places where emphasis should be placed or removed so as to clarify ideas within a broad context. Hence the discourse analysis aids in assigning prosodic features to words and phrases, and in choosing the most appropriate intonation and duration contours for the speech.

1) TTS System Performance [17]: The only proper way to judge TTS performance is to listen to one or more paragraphs of speech produced by the system. Without this capability, the next best way of describing TTS performance is in terms of intelligibility scores and MOS scores of quality. The best TTS systems achieve word intelligibility scores of close to 97% (natural speech achieves 99% scores); hence the intelligibility of the best TTS systems approaches that of natural speech. MOS scores for the best TTS systems are in the 3.0–3.5 range, indicating that the current quality is judged to be in the fair-to-good range. The computation necessary to support full TTS systems is modest by today's standards (2 MIPS processing), with somewhere between 2 and 6 Mbytes (MB) of memory required for units, dictionaries, rules, and program code.

2) Requirements for Improved TTS [10], [12]: In order to improve the quality (naturalness) of current TTS systems, three areas must be addressed. These include:

1) **Improved model of source-filter interactions:** The current model, which assumes independence between the vocal tract source excitation, and the vocal tract filter, is grossly inadequate—especially when trying to model female speech. A more realistic model, possibly incorporating nonplane wave propagation in the vocal tract, is required, along with improved understanding of how the source rate of periodicity influences the vocal tract shapes, for female talkers, so as to transmit the most sound energy through the vocal tract.
2) **Improved prosody rules:** Experiments have shown that when natural duration and pitch are "copied" onto a TTS utterance, while preserving the sound units that the TTS system generates from the text, the quality of the resulting synthetic speech improves dramatically. Hence it is mandatory to develop better rules for generating duration and pitch contours for utterances.
3) **Improved linguistic analyses:** Although current linguistic analyses have provided significant improvements in naturalness of TTS systems, they still have a long way to go before the system sounds like it "knows what it is talking about." Until such understanding of what it is saying is achieved, TTS systems will sound choppy and "unsure of themselves" over time.

E. TTS Applications

In spite of the synthetic quality of current TTS systems, a number of interesting and important applications of TTS have evolved and are currently in use. These include the following:

1) Network voice server which provides access to either e-mail or FAX via synthetic speech. Clearly, this service is invaluable to people "on the go" who have no direct access to alternative communications services, e.g., terminals, FAX machines.
2) Voice previewer for draft material which provides an alternative medium (to reading) to spot errors in text, determine improper constructions, and, in general, to get a feeling for the message contained in the written material.
3) Information about course availability which provides students with an opportunity to "hear" more about potential courses than can be included in a standard catalog, or to provide "up-to-date" course information.
4) Feedback on installation and repair of telecommunications equipment which provides service people with direct feedback on special requirements, checks out the equipment, and provides straightforward verification that the servicing was done correctly.
5) Wakeup/reminder call services for hotels which provide individualized messages for travelers to remind them of appointments, schedules, or just to get them out of bed in the morning.
6) Automated order inquiry and status which enables both salespeople and customers to track orders from inception to delivery. This service is especially sensitive to proper name pronunciation as it is personalized to individual salespeople and customers, as well as to products which often have distinctive names associated with them.
7) Course registration which enables students at universities to compete for available courses on a fair basis (i.e., without waiting in endless registration lines) and to modify schedules easily in response to unavailable courses and conflicts in schedules.
8) Directory assistance (including addresses) which enables customers to access directories of names and addresses directly without going through the delay or expense associated with an attendant. Again this application is especially sensitive to accurate pronunciation of both proper names and street addresses.
9) Business locator service which enables customers to find the nearest location of a business or service without the help of an attendant and without the need to call the business directly. This application could also be coupled with a direction finder to enable the customer to determine the "best" way to travel to the location provided by the service.
10) Reverse directory assistance providing the customer with the name and address associated with a specified telephone number so as to allow customers to screen incoming calls in order to decide which ones to answer directly, and which ones to defer to some type of messaging service.
11) Banking services providing the customer with access to and control of bank accounts including account status, check status, and bill paying options.

12) HELP service lines for medical and legal services, documentation on equipment, and for help in getting repairs done, either by the service person or for the home hobbyist.
13) Dual party relay service which is a federally mandated service to help speech and/or hearing impaired customers avail themselves of telecommunications services. We discuss this particular application of TTS more thoroughly in the next section.

1) Dual Party Relay Service: A key application area for TTS is the dual party relay (DPR) service which is an 800 service and provides a means for speech and/or hearing impaired customers to "hold a conversation" with a nonimpaired customer using a device called a Telecommunications Device for the Deaf (TDD). The TDD is a keyboard which allows the impaired user to generate text sequences which are transmitted to an attendant who speaks the text to the nonimpaired user. Similarly the attendant "translates" the speech from the nonimpaired user to a text sequence which is transmitted to the impaired user. The disadvantages of this normal DPR service include: lack of privacy, as the attendant has direct access to both sides of the conversation; cost, since the attendant is tied up for the duration of the call; speed, since the attendant has to see a sufficient amount of the text to speak it, and a sufficient amount of the speech to convert it to text properly; and, potentially, accuracy as the attendant must follow both the text and the speech without often understanding the context of the conversation.

The way in which TTS is utilized is to provide a direct link between the impaired user and the hearing party by translating the TDD text directly to speech. In this manner there is potentially great cost savings as an attendant is not needed for the majority of time of the call (only for translating from speech to text), and greatly improved privacy since no single attendant "hears" more than a small fraction of any call.

The major problems associated with DPR using TTS are related with the freely generated TDD texts, whereby customers use constructions that are specialized to the application, along with nongrammatical inputs, a lack of punctuation, and a high frequency of improperly spelled words. Hence a typical TDD sentence might be:

"When Do I Will Call Back U Q Ga"

where the "U" is short for "you", the "Q" is a question indication, and the "Ga" is the command to "Go ahead", i.e., respond to the text. The TTS system must not only handle these unusual, ungrammatical, constructions, it must do so in near real time. Hence it cannot process the entire TDD text sequence to detect and correct errors but must do so in an almost serial manner as the text is received. Further, it must determine syntactic breaks sequentially, without knowledge of what text follows. These issues are difficult to handle with clear text; with DPR text they are even more difficult to get right.

By way of example, the following paragraph illustrates the problems associated with linguistic analysis of DPR text:

"we are real very enjoying with the CRS thats what they are very good service and very good impression to us as they are very manner to do good for us and they always accept us for call for us as good respectful what we want to tell something and important and many things as we need to have"

The ability to parse such text and provide appropriate, and correct, pauses is crucial for this application.

V. Speech Recognition [18]–[26]

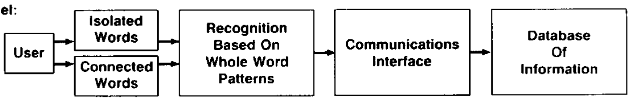
The goal of speech recognition is to provide enhanced access to machines via voice commands. The idea of "enhanced" access is a key one since, for most applications, there are viable alternatives to voice control, including keyboards, touch panels, mice, etc. Thus for voice technology to be of value means that the voice interface to the machine has to be a natural one in which voice input is a reasonable way of requesting information, and the interface performs reliably (with high accuracy) and robustly for all users and in all environments. Figure 27 shows a block diagram of a communications model of voice access to and control of a machine. In order to access a database of information, the user is assumed to speak commands in the form of either an isolated word sequence or as a sequence of words drawn from a small vocabulary (e.g., digits). We refer to this second form of spoken input as connected word sequences. Recognition of the spoken input is based on whole-word patterns; hence, the output of the recognizer is either the appropriate command word, or a recognized sequence of words. A communications interface is used to access the database for the appropriate information, which is transmitted back to the user using TTS messages.

There are a wide range of factors that influence performance of speech recognition systems, including the following:

1) **Speaking format:** There are three standard modes of speaking to a machine; namely, isolated word (phrase) mode, connected word mode, and continuous speech mode. Isolated word (phrase) recognition is primarily used for so-called "command and control" tasks where the machine responds appropriately to each spoken command. Connected word recognition uses fluent speech with highly constrained vocabularies and is used for tasks such as order entry, credit card validation, and digit dialing, where sequences of words (e.g., digits) specify the information being sought by the machine in order to enable completion of some transaction. Finally, continuous speech recognition is used for dialog sessions with the machine in order to perform tasks such as database management and access, voice dictation, and language translation.
2) **Degree of speaker dependence:** This factor describes whether the recognition system needs to be trained to the speech patterns of individual users (so-called speaker-dependent [SD] systems), or can work reliably with users who have never (or seldom) seen

Goal: To Provide Enhanced Access To Machines Via Voice Commands

Model:



Factors:

1. Word Format - Isolated Words/Phrases
   - Fluent Speech
2. Speaker Trained Or Speaker Independent
3. Vocabulary Size And Complexity
   - Small, Low Complexity (Digits)
   - Large, High Complexity (5000 Most Common Words)
4. Task To Which Applied - Semantic, Syntactic Constraints
5. Cost, Method Of Implementation

Fig. 27. Communications model of database retrieval system using speech recognition to transmit database requests for information.

or used the system (so-called speaker independent [SI] systems).

3) **Vocabulary size and complexity:** This factor describes the range of vocabulary words and phrases which the system understands. There are many useful and interesting tasks requiring small to moderate size vocabularies (e.g., digit strings, simple commands from menus); however, ultimately systems must be able to reliably recognize upwards of 50 000 words for tasks such as voice dictation of letters, natural language access to databases, etc.
4) **Task constraints:** As the recognition vocabulary size grows, the number of possible combinations of words to be recognized can grow exponentially. Hence, some form of task constraint, in the form of formal syntax (defining which words can follow other words in different contexts) and formal semantics (defining which sentences make sense for the current status of the task transaction) is required to make the recognition task more manageable, by reducing the number of possible word candidates at any point in the utterance, and thereby providing higher recognition task accuracy.
5) **Cost, method of implementation:** Speech recognition by machine is often computationally quite expensive (upwards of 1 gigaflop/s is required for real-time operation for some problems). Hence, a limiting factor is often what can be done with reasonable, but limited, computational resources.

A. General Applications of Speech Recognition

Applications of speech recognition technology fall into two broad areas; namely, telecommunications and business applications. In the telecommunications area, some representative applications include:

1) **Expanded use of rotary phone for menu-based (IVR) services:** Such services currently are unavailable without a touch-tone phone. In addition, even for users with access to touch-tone phones, a voice recognition interface can be more attractive than the standard button-pushing alternative because the service names are spoken rather than having to push buttons associated with the service. Thus for access to different parts of a department store, it is more natural to speak the words "hardware" or "furniture" rather than to remember to push Button 3 for "hardware" or Button 5 for "furniture."
2) **Repertory dialing:** Voice dialing of telephone numbers and names provides the opportunity for hands-free, eyes-free control and use of a telephone. This is especially important for mobile telephony when the eyes and hands are usually tied up with the process of driving and controlling an automobile.
3) **Catalog ordering:** Most catalogs consist of letter and number codes attributed to each item in the catalog, often with a great deal of redundancy built into the codes. Hence, ordering items from a catalog, by voice, is a natural way of interacting with the database of items associated with the catalog.

In the business area, some representative applications include:

1) **Data entry for filling out forms:** Such applications are highly repetitive and generally are performed by a small staff of people who can afford to train the system to recognize individual word patterns. Typically, vocabularies for this application are small to moderate in size (e.g., from 10 to 200 words).
2) **Keyboard replacement or expansion:** Here the recognition task is to replace sequences of keystrokes with a single voice command (a voice macro) or to replace the keyboard entirely using spoken input.
3) **Database access:** The recognition task is to query a database to determine specific information contained within the database. Hence, an airline's reservation system could be queried to determine available flights between specified cities, flight costs, type of aircraft, etc.
4) **Test, inspection, and process control in manufacturing:** Here the recognition task provides eyes-free, hands-free, access to monitoring any step in the manufacturing process so as to detect defects, production problems etc.,

B. Speech Recognition Technology [22], [24], [25]

Before reviewing some basic techniques for speech recognition by machine, it is worthwhile discussing the

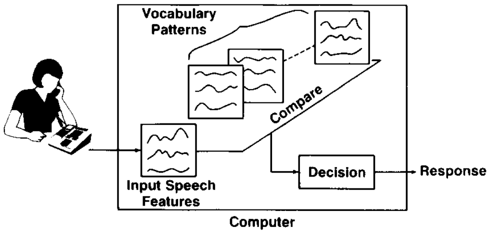


Fig. 28. Simple block diagram of pattern-recognition model for word recognition.

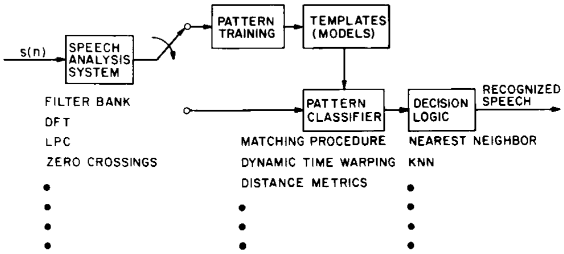


Fig. 29. Statistical pattern recognition model applied to speech recognition.

question as to why speech recognition is so difficult. Although there is no simple answer here, perhaps the most important factor which limits performance of various systems is variability. This variability comes in different forms including:

1) variability of sounds (e.g. words, phrases, subword units) both within a single speaker and across speakers;
2) transducer/channel variability including differences in signal characteristics due to the use of carbon button and electret microphones, speakerphones, and cellular phones;
3) background noise variability from extraneous speech (e.g., TV, radio, side conversations) or from transient acoustic events such as road noise, door slams, etc.;
4) speech production variability including mouth clicks, breath noise, hesitations in speaking, and extraneous speech.

A key factor is that the sources of variability cannot generally be eliminated; hence they must be modeled directly in the speech recognition technology.

Based on the above, there are three things a speech recognizer must handle properly, and these include:

1) Speech detection, i.e., the separation of speech from the background so that recognition is performed only on speech input provided by the user.
2) Recognition of the spoken input, based on pattern recognition technology (including both deterministic and statistical methods), or on acoustic–phonetic methods, or on neural network methods.
3) Human factors, properly accounting for the presence of extraneous speech, associated "uhm's" and "ahs," and cases where the user backed up and started over.

Although several approaches to speech recognition have been proposed, the most popular (and successful) approach has been one based on standard pattern recognition technology, as illustrated in Fig. 28. Basically, the system uses a set of word and/or phrase patterns created using a pattern training program. These patterns can be typical spectral patterns of words, averages of spectral patterns of words across different talkers, or sophisticated statistical models that include spectral mean and spectral variance statistics derived over the time duration of the word.

The way in which isolated word recognition is carried out once the vocabulary patterns have been created and stored is to record the input speech features (called the unknown spectral pattern) and to compare them against each of the stored vocabulary patterns. The pattern that best matches the input speech features is determined, and, if the match is close enough, the decision box provides a response consistent with the recognized word. If the match is not sufficiently close, no decision is made and the user either can repeat the word or choose an alternative way of making the request to the system.

The pattern recognition system of Fig. 28 handles only the problem of recognizing spoken input. We will see later on how to handle speech detection and human factors within the framework of the pattern recognition approach.

1) The Statistical Pattern Recognition Model: Figure 29 shows a more detailed block diagram of the statistical pattern recognition model for speech recognition. The speech signal $s(n)$ is first analyzed into a set of short-time parameters which characterize the time-varying nature of the signal. These parameters could be spectral parameters, such as the output signals from a filter bank, a DFT, or an LPC analysis, or they could be temporal parameters, such as the locations of various zero or level crossings times in the speech signal.

The pattern training block creates either templates (average characterizations of the speech parameters for a given word or phrase) or statistical models (characterizations of both the mean and variance of the speech parameters—usually according to a particular statistical model). The pattern training algorithm for templates is generally a clustering procedure which tries to cluster parameter sets from multiple versions of a word or phrase into consistent groups (clusters) so that the average intracluster distance between word tokens is significantly smaller than the average intercluster distance. Figure 30 illustrates this point by showing each word token as a dot in a simple parameter space (highly stylized). Groups of word tokens (dots) are joined together in clusters, e.g., $C1$ through $C5$, where each cluster ultimately is used to create an individual word template. Also shown in Fig. 30 are so-called outlier word tokens, $O1$ through $O8$, which are too far from any existing cluster and therefore are discarded in the pattern training procedure. For statistical models, a segmental K-means variant on the clustering procedure is used to create the models.

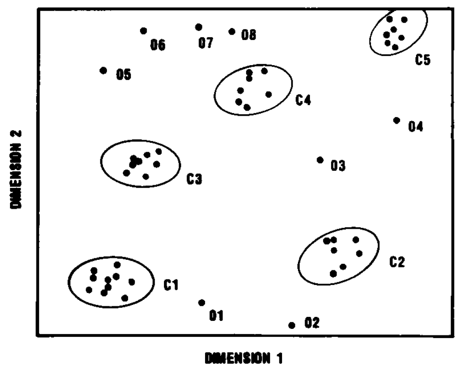


Fig. 30. Example illustrating clustering of word tokens into well-defined clusters ($C1$ to $C5$) with outlier tokens ($O1$ to $O8$).

The pattern classifier has the job of comparing the speech pattern from the unknown spoken word (or phrase) with the stored patterns (the templates or models) and generating a dissimilarity or distance score for each such comparison. One key problem that arises when comparing speech patterns is that of time normalization, as illustrated in Figs. 31 and 32. Figure 31 (the left side) shows the log energy contour of two patterns (for the spoken word / seven/)—called the reference (either template or model) and the test (the unknown input). It can be seen that the inherent duration of the two patterns, 30 and 35 frames (where each frame is a 15-ms chunk of speech), is different, and that linear alignment is grossly inadequate for aligning events within the two patterns (compare the locations of the vowel peaks of the two patterns). Hence a procedure, called dynamic time warping, is used to nonlinearly align the time scales of the reference and test patterns via an alignment path which is optionally determined using a dynamic programming algorithm. The results of the dynamic time warping are shown in Fig. 32 where we see, at the top, the linear alignment of the patterns, and, at the bottom, the nonlinear alignment. It is clear that the nonlinear alignment is significantly better than the linear alignment and provides a more realistic measure of dissimilarity or distance between the patterns.

The final block in Fig. 29 is the decision logic which finds the closest match to the unknown pattern and decides if the quality of the match is good enough to make a recognition decision. If not, the user is asked to provide another token of the word (or phrase) and the process repeats itself.

2) Hidden Markov Models [18], [19], [21]: The most popular statistical model used in speech recognition is the hidden Markov model (HMM). For this procedure, training consists of estimating the parameters (means and covariances) of a probabilistic model for each word. To classify an unknown utterance, one computes the likelihood that it was generated by each of the models derived during training. The utterance is (generally) recognized as the word whose model gives the highest likelihood score.

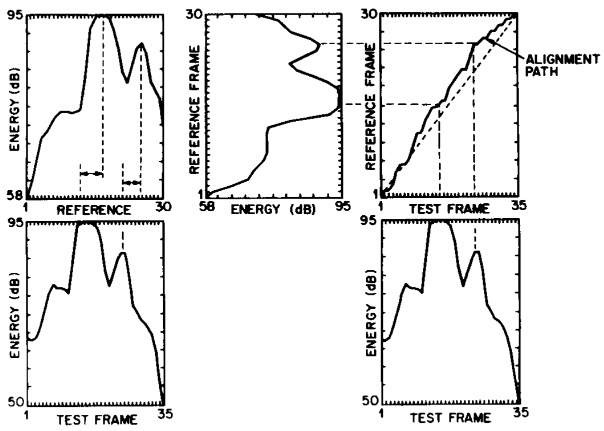


Fig. 31. The problem of time aligning a pair of word utterances. The left side of the illustration shows energy contours of the two words being time-aligned and the right side shows the time-alignment process.

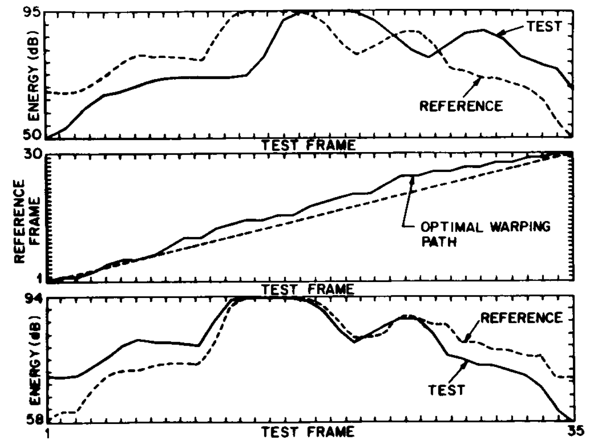


Fig. 32. Results of time aligning two versions of the word "seven," showing linear alignment of the two utterances (top panel); optimal time-alignment path (middle panel); and nonlinearly aligned patterns (lower panel).

Figure 33 shows the structure of a standard left-to-right HMM where the word is represented using a 5-state model. Within each state of the model, the speech parameters are characterized by several density functions which are generally some type of mixture of Gaussian densities. Hence a set of spectral parameters would be characterized by a set of means, covariances, and weights of M Gaussian mixtures in an observation density $\boldsymbol{b}(O)$, whereas log energy would be characterized by a separate density, as would state duration. Transitions between states of the model are again statistically characterized by a state transition matrix whose parameters are estimated as part of the training procedure.

3) Connected Word Recognition [20], [23]: It is relatively simple and straightforward to extend the word recognition model of Fig. 29 to handle word sequences. The basic idea is illustrated in Fig. 34 which shows an unknown word sequence and a group of word reference patterns which constitute the vocabulary for the recognizer.

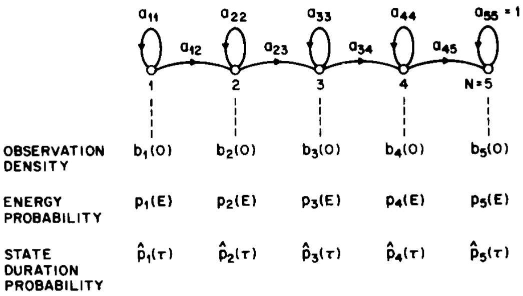

a_{11}
a_{22}
a_{33}
a_{44}
$a_{55} = 1$
a_{12}
a_{23}
a_{34}
a_{45}
1
2
3
4
N=5
OBSERVATION DENSITY
$b_1(O)$ $b_2(O)$ $b_3(O)$ $b_4(O)$ $b_5(O)$
ENERGY PROBABILITY
$p_1(E)$ $p_2(E)$ $p_3(E)$ $p_4(E)$ $p_5(E)$
STATE DURATION PROBABILITY
$\hat{p}_1(\tau)$ $\hat{p}_2(\tau)$ $\hat{p}_3(\tau)$ $\hat{p}_4(\tau)$ $\hat{p}_5(\tau)$

Fig. 33. Characterization of a word (or phase, or subword) using an $N(5)$ state, left-to-right, hjdden Markov model, with continuous observation densities in each state of the model.

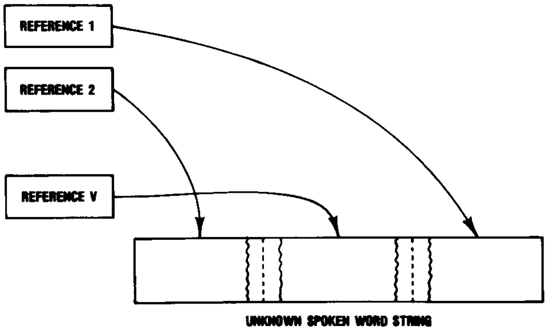


Fig. 34. Illustration of the problem of matching a connected word string, spoken fluently, using whole word patterns concatenated together to provide the best match.

The basic idea is that we would like to concatenate word reference patterns in every possible combination of $1, 2, 3, \cdots, L$ patterns, match these concatenated sequences to the unknown word sequence, and choose the concatenated sequence that best matches the spoken word string. Although such an exhaustive algorithm appears to grow in computation exponentially with the number of words in the string, algorithms have been devised which limit the growth in computation to be linear with the length of the word string. Hence, with some modification, the basic speech recognition algorithm of Fig. 29 can be used to handle connected word sequences.

One could also introduce the concept of a word grammar into the recognizer, where the grammar serves to restrict the possible word sequences that are possible for recognition. For example, a digits grammar could restrict the search to only 7- or 10-digit strings for telephone numbers, or to beginning with a known prefix code for credit card numbers. It is relatively straightforward to integrate such grammars (when represented as finite-state networks) into the recognition procedure.

Using the concept of a word grammar with the connected word recognizer, the problem of segmenting a signal into regions of speech and background noise becomes one of recognizing the entire recording interval as a sequence of "background"—connected words—"background," which is readily handled by the system. Thus the only change required is to create a reference pattern for the "background," and then treat it the same as any other vocabulary pattern, within the grammatical constraints of the system.

4) Continuous Speech Recognition [18], [19], [26]: The ultimate goal of speech recognition is to be able to recognize fluent (continuous) speech with a vocabulary that is essentially unlimited. Although we have not yet succeeded in reaching this goal, systems for continuous speech recognition with vocabularies on the order of 1000–20 000 words currently do exist and work reasonably well in the research laboratory.

A block diagram of a typical continuous speech recognition system is shown in Fig. 35. The first two blocks of the system; namely, feature analysis and unit matching, are essentially the same ones used in word or connected word recognition (assuming the recognition units are words). The key difference occurs when the units become subword units, i.e., when pieces of words, e.g., phonemes, are used instead of whole words. This leads to the concept of representing words in terms of a word dictionary (a lexicon), where each word has one or more decompositions into basic subword units, and using lexical decoding to match words in the speech. The final two blocks, namely, syntactic analysis based on a word grammar, and semantic analysis, based on a task specification, are used to restrict the sequence of words that needs to be matched against the input speech. Since the word vocabularies are generally large, the restrictions of the grammar and syntax are needed to keep the matching procedure computationally feasible for most large vocabulary tasks.

Generally, the last 4 blocks of Fig. 35, namely, unit matching, lexical decoding, syntactic analysis, and semantic analysis, are all integrated together into a large finite-state network which is searched efficiently, using a beam search procedure, to give the best sentence interpretation of the spoken input. In this manner, modern workstations can handle vocabularies of up to 20 000 words in near real-time.

5) Speech Recognition Issues: There are many unresolved issues in speech recognition that severely limit its utility in practical applications. These include the following:

1) **Handling speech that is not in the recognition vocabulary.** This problem is pervasive in that it exists for all speech recognition modes. Hence, even for an isolated word recognizer, users will often say "yes, please" when asked to respond "yes" or "no" to a question. For large vocabulary recognition, the issue is compounded since most users will not know the vocabulary and therefore will be unaware they are using words that the recognizer does not understand.
2) **Recognizing speech in noisy environments.** This problem is a key one in automobiles (e.g., for voice control of cellular phones), in airplanes, train stations, etc., where the noise level is high, and little can be done to reduce the level. Robust recognition techniques are required to handle these adverse conditions for recognition.
3) **Adaptive training.** Although speech reference patterns are, in theory, derived through a training pro-

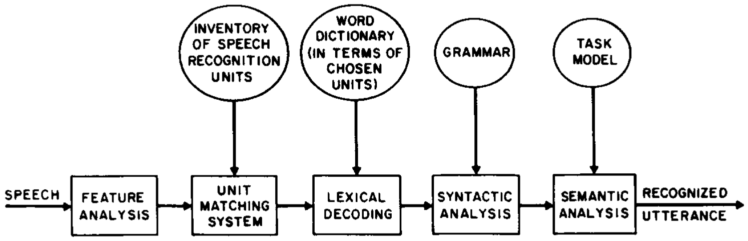


Fig. 35. Block diagram of a large vocabulary speech recognition system (bottom-up approach) incorporating syntactic and demantic analysis modules.

Table 2 Word Error Rates for a Range of Speech Recognition Systems

Technology	Task	Syntax	Mode	Vocabulary	Word Error Rate (%)
Isolated Words	none	none	SD	10 digits	0
				39 alphadigits	4.5
				1109 basic English	4.3
			SI	10 digits	0.1
				39 alphadigits	7.0
				129 airline words	2.9
Connected Words	digit strings	known-length string	SD	10 digits	0.1
			SI	11 digits	0.2
	airline reservations	finite state grammar (perplexity = 4)	SD	129 airline words	0.1
Fluent Speech	Naval Resource Management	finite state grammar (perplexity = 60)	SI	991 words	4.5
	ATIS	finite state grammar (perplexity = 12)	SI	1800 words	4.0
	Wall Street Journal	finite state grammar (perplexity = 200)	SI	20 000 words	13.0

cedure and thereafter unmodified, in practice the performance of the recognizer will degrade if the conditions in which it is used differ significantly from the conditions in which it is trained. Adaptive training is capable of modifying and updating the reference patterns to track such differences, and thereby improve recognizer performance.

C. Speech Recognition Performance

A summary of the performance of speech recognizers, based on laboratory evaluations, for the three technology areas (isolated words, connected words, fluent speech), and for different task applications, is shown in Table 2. (The reader should note that real world performance of most recognition systems is significantly worse than that of the laboratory evaluations shown in Table 2.) The measure of recognizer performance is the word error rate (in percent) for a given vocabulary, task, and syntax (grammar).

For isolated word recognition, the results are given without any task or syntax constraint—i.e., every word in the vocabulary is assumed equally likely. For a digits vocabulary the word error rates are quite low both in SD (speaker dependent) mode (0%) and in SI (speaker independent) mode (0.1%). For an alphadigits vocabulary, consisting of the spoken letters of the alphabet, the digits, and three command words, all spoken over dialed-up telephone lines, word error rates are 4.5% for SD mode and 7.0% for SI mode. Considering the confusability among spoken letters (e.g., B, C, D, E, G, P, T, V, Z), these results are actually quite impressive for telephone-bandwidth speech. For more distinctive vocabularies, word error rates are quite reasonable at 2.9% for 129 airline words (SI), and 4.3% for 1109 basic English words (SD).

For connected word recognition, word error rates for known length digit strings are again quite low at 0.1% (SD) and 0.2% (SI). Similarly, for an airline reservations task, with a grammar whose perplexity (average word branching factor) is low (4), the word error rate in SD mode is 0.1%. For fluent speech recognition, results are based on DARPA funded research on 3 tasks; namely, a ships database task (Naval Resource Management), an airline travel task

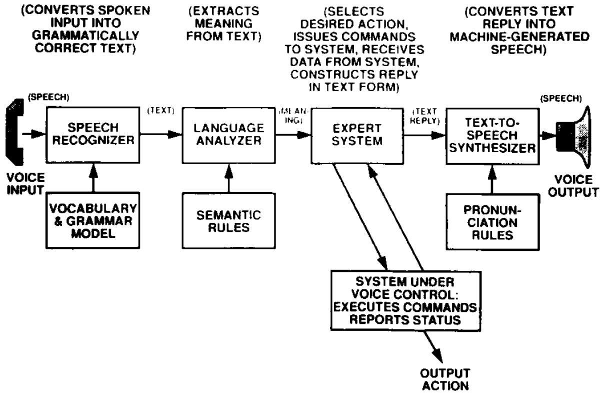

(CONVERTS SPOKEN INPUT INTO GRAMMATICALLY CORRECT TEXT)
(EXTRACTS MEANING FROM TEXT)
(SELECTS DESIRED ACTION, ISSUES COMMANDS TO SYSTEM, RECEIVES DATA FROM SYSTEM, CONSTRUCTS REPLY IN TEXT FORM)
(CONVERTS TEXT REPLY INTO MACHINE-GENERATED SPEECH)
(SPEECH)
SPEECH RECOGNIZER
(TEXT)
LANGUAGE ANALYZER
(MEANING)
EXPERT SYSTEM
(TEXT REPLY)
TEXT-TO-SPEECH SYNTHESIZER
(SPEECH)
VOICE INPUT
VOCABULARY & GRAMMAR MODEL
SEMANTIC RULES
PRONUN-CIATION RULES
VOICE OUTPUT
SYSTEM UNDER VOICE CONTROL: EXECUTES COMMANDS REPORTS STATUS
OUTPUT ACTION

Fig. 36. Model of a task-specific voice control and dialog system.

(ATIS), and speech read from the *Wall Street Journal*. The vocabulary sizes and grammar perplexities of these 3 tasks are 991 words (perplexity 60), 1800 words (perplexity 12), and 20 000 words (perplexity 200), with laboratory evaluation word error rates of 4.5%, 4.0%, and 13.0%, respectively. Although these word error rates are quite impressive, it should be stressed that sentence error rates, for these tasks, are significantly higher. Hence this technology is not yet suitable for use in real-world applications.

D. Speech Recognition Applications

If one considers speech recognition applications in the broad concept of a task, as shown in Fig. 36, we see that the recognizer is only a part of the overall transaction. Thus the entire process consists of speech recognition, which converts the spoken input into grammatically correct text, language analysis, which extracts the meaning from the text, an expert system, which selects the desired action, issues commands to the system which is under voice control, receives data from the system, and constructs a reply in text form, and finally a text-to-speech synthesizer, which converts the text reply into machine-generated speech which is sent to the user. Hence the overall system is both a dialog system (as far as the user is concerned), and a control system which carries out the action requested via the voice command.

Based on the task-specific model, there is a broad range of applications of speech recognition both within telecommunications and in the business arena. To understand how speech recognition can be applied effectively to different problems, we must first understand the requirements that the proposed task must satisfy. These include the following:

1) **actual benefit to a user:** The use of voice recognition for control should be natural and of value to the user. It cannot be a novelty to attract attention or to temporarily increase sales.
2) **user friendly:** The recognizer should be easy to use and the commands should be mnemonic. The system should be robust to the ways in which users interact with the system.
3) **accurate:** The system must achieve a specified level of performance (e.g., word accuracy greater than 95%), so that the user is motivated to continue using the system.
4) **real-time response:** It is mandatory that users be provided a discernable system response in a timely manner, much as they see characters echoed back on a terminal when they type, or hear audible ringing after dialing a sequence of telephone digits. Such feedback is mandatory so that users feel in control of the actions of the system.

Even with the above restrictions, there are several other characteristics that the task must possess for the use of voice recognition to be successful. These include:

1) **fail soft application:** Speech recognizers are error prone devices — they are guaranteed to make mistakes some percentage of the time. Fail soft applications are those that can tolerate errors, no matter what type — i.e. those where the cost of a recognition error is low. An example of this type of system is a menu-based system where an error leads to an incorrect menu item. Just as with a mouse type system,when a menu error is made, the user can cancel the menu and start over, or, if possible, back up to the preceding menu and then correct the mistake.
2) **self-detection/correction of errors:** An alternative to making errors is to provide mechanisms in the recognition task to reduce the error rate through the use of check digits (for credit cards), list syntax (for name directories), and multiple candidate strings (where a more detailed search can be made among the candidate strings).
3) **verification before proceeding:** By using information provided by the recognizer (e.g., distance scores, scores for multiple candidates), the system can automatically detect potential errors and ask the user to confirm the recognition before carrying out the command. In this manner, whenever the recognizer is not confident of its decision, the user is asked to aid in the error detection and correction process.

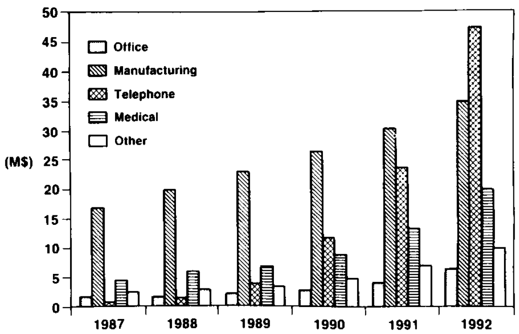


Fig. 37. A plot of the hardware sales within the total US speech recognition market, segmented into five general areas, for the years 1987–1992.

4) **rejection/pass on to attendant:** As opposed to the preceding procedure, the task can automatically reject some percentage of the input speech, and pass on the spoken input (generally recorded digitally and saved temporarily) to an attendant who listens to the speech and either confirms or corrects the recognizer decision.

Finally, there are a set of technology requirements for speech recognition that provide the basis for a range of successful applications. These include:

1) word spotting capability, namely the ability to recognize either a command word or a command sequence within fluent speech;
2) barge-in, namely the ability of the user to speak over the voice prompt (thereby canceling the prompt) and be recognized correctly: this feature is invaluable for experienced users who do not need to listen to the voice prompt to know what to say to the system;
3) robustness, namely the ability to maintain a consistent level of performance for different users, backgrounds, handsets, and communications channels;
4) rapid deployment, namely the ability to create new services without the need for extensive vocabulary training.

1) Speech Recognition Markets: Figure 37 shows a plot of sales of speech recognition hardware in the US from 1987 to 1992, broken down into five market segments. Although the manufacturing segment started at a high level of sales in 1987 (relative to the other segments), by 1992 the exponential growth in sales in the telecommunications area overtook manufacturing, and is currently the largest market segment in speech recognition. Although this figure shows hardware sales, it is interesting to note that the service revenue associated with speech recognition technology is often one or two orders of magnitude larger than the hardware revenue. Hence the total impact of speech recognition, in the marketplace, is significantly larger than what is implied by the numbers in Fig. 37.

2) Speech Recognition Applications to Telecommunications: There are two broad categories of speech recognition applications in telecommunications; namely, those which provide cost reduction, and those which generate revenue. Cost reduction applications are primarily those which replace human attendants by speech recognition devices. For these applications the accuracy and the efficiency of the recognizer is of paramount concern, since the tasks being performed by machine were previously performed by live attendants. The benefit of these applications is that large cost savings can be achieved. The limitation is that since the cost savings go to the service provider, the customers may not be cooperative or forgiving of the technology limitations. Their perception could be that the technology has degraded, rather than improved, the service received. Hence it is critical that such cost reduction applications be carefully chosen.

The second broad category is those applications that generate revenue. In general, such applications provide a service or a capability that was previously not available (often because it would have been too expensive to provide the service using human attendants). Hence, in this case, since the benefit is to provide user access to services or information that was previously not possible, the customers are generally cooperative, and quite forgiving of technology limitations.

3) Cost Reduction Applications: Examples of telecommunications services which provide cost reductions include the following:

1) automation of operator services, including the AT&T VRCP (Voice Recognition Call Processing) Service for automation of 0+ calls, and the Bell Northern AABS (Automated Alternative Billing Service) for automation of the response to accepting charges for collect calls;
2) automation of directory assistance, including front end processors for determining the city name by Nynex and Bell Northern, and full directory listing retrieval based on either spelled or spoken names;
3) voice dialing services, either by name (the so-called alias dialing), or by number (direct dialing).

By way of example, consider the VRCP system introduced by AT&T in 1992. VRCP is a menu-based service for automating many types of billing functions in operator assisted (0+) calls. This application required a vocabulary of only five words, namely, "collect," "person-to-person," "third number," "calling card," and "operator," corresponding to the five types of calls that can be placed. The *a priori* statistics on usage show that about 50% of operator-assisted calls are collect calls, 12% are billing to a third number, 1% are person-to-person, 10% are calling card (with operator entry of card number), and 27% are inquiries for various types of assistance. The system is inherently a speaker-independent system with telephone input.

Preliminary experience with the system indicated that word spotting was essential for this service, as about 20% of the calls used commands of the form "*collect* call please," "I'd like to place a *calling card* call," "*Person-to-person* from Tom to Alice." In addition, since there were so many repeat users of the service, barge-in was found to be a necessary feature for these experienced users who were accustomed to interrupting the voice prompt with touch-tone signals, and wanted to do the same with voice. Finally, it was found that effective voice prompts were essential to enable users to effectively use the service the first time. We will return to issues related to "ease of use" later in this section.

To realize the impact of this voice recognition service, consider the usage statistics. In 1992 AT&T averaged about 145 000 000 calls per day on the 0+ network. About 92% of these calls were calling card calls placed automatically by users keying in both the called number and the calling card number using the touch-tone receiver. The remaining 8%, or about 11.6 million calls, were handled by operators. In full depolyment, VRCP will handle about 4.2 billion calls a year, at a cost savings on the order of $300 M annually. The cost savings on this one simple application are an order of magnitude larger than the total hardware sales for telecommunications-based speech recognizers.

Another interesting cost reduction application is validation of credit card sales for companies like American Express or VISA. This service is used by merchants who do not use the modem dialers to validate credit card sales, but instead manually dial into an 800 number and normally speak to an attendant. Instead of the attendant, a voice recognition system prompts the merchant to enter a 10-digit merchant ID number, a 15-digit credit card number, and the dollar value of the transaction. Both the merchant ID number and the credit card number are highly constrained sequences; hence accurate recognition is not a major problem, even over dialed-up telephone lines. However, accurate recognition of the transaction dollar amount is problematic, since there is no "standard" way of speaking this amount. For example, a transaction of $116.33 could be spoken in a number of ways, including:

"one sixteen and thirty three"
"one hundred sixteen point thirty three"
"one one six dot three three"
"one sixteen dollars thirty three cents"

Such sequences are called natural number sequences and are much more difficult to recognize accurately than connected digit strings because of the lack of regularity, the confusability between words like /six/, /sixteen/, and /sixty/, and the difficulty of deciding which part is the dollar amount, and which part is the cents of the transaction.

4) Revenue Generation Applications: Examples of telecommunications services which generate revenue and provide new capabilities include:

1) voice banking services, such as the NTT ANSER system;
2) voice prompter service, consisting of touch tone replacement by voice recognition, as introduced by AT&T in its Intelligent Network;
3) information access systems, such as the Northern Telecom stock price quotation system;
4) directory assistance call completion, whereby the system actually dials the call based on recognizing the spoken response provided by the directory services provider: such services are available from NYNEX and AT&T;
5) reverse directory assistance, whereby a customer can retrieve a name and address associated with a given telephone number: this service is available from NYNEX, Bellcore, and Ameritech.
6) information services, such as sports scores, traffic reports, weather reports, theatre bookings, etc.

Consider the NTT ANSER system for voice banking. This service is a menu-based system for providing banking information and services to customers. (Previously there was no way for customers in Japan to access this information.) In particular, customers could query the system to determine account balances, to make bill payment queries, and to obtain credit information. The system is speaker-independent, isolated word, with telephone input, and a vocabulary of 16 words consisting of the digits plus 6 command words. A typical transaction is of the form:

Customer	System
(Calls Center)	"Hello, this is the NTT bank telephone service center. What is your service number?"
"one, one"	"You are asking for your account balance. What is your branch number?"
"one, two, ..."	"What is your account number?"
"three, four, ..."	"What is your secret number?"
"five, six, ..."	"Your current balance is 153 000 yen. If you would like to have your balance repeated, please say "Once More". If not, say 'OK'."
"OK"	"Thank you very much"

At the end of 1992 the ANSER system was handling 360 M calls per year, and providing about $30 M in revenue per year. Another good example of a revenue generation system is the Voice Prompter Service for call distribution via voice commands—i.e., replacing the touch tone queries with voice queries. Typical usage of this service would be transportation, e.g., AMTRAK, where the user could choose among "Departures (1)," "Arrivals (2)," "Reservations (3)," and "Special Services—e.g., Metroliner (4)," by either speaking the commands, or the numbers associated with the commands. For hotels, the user could choose among "Guest Rooms (1)," "Reservations (2)," "Hotel Operator (3)," etc.

Figure 38 illustrates an information service built by Telefonica in Spain based on the Voice Prompter Service in

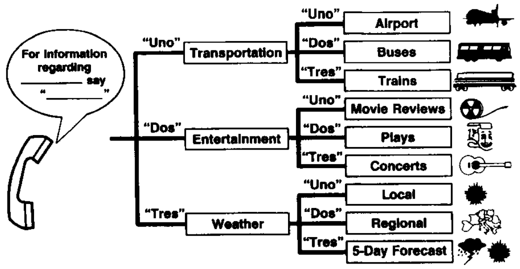


Fig. 38. Illustration of the use of a Voice Prompter Service for accessing information databases using Spanish commands.

Spanish, using the Spanish digits "uno," "dos," and "tres." Based on this menu system, the user could get information on any of nine topics with just two voice commands.

A final example of a revenue generation service is the Voice Interactive Phone (VIP) service introduced by AT&T. The service allows customers to access a wide range of telecommunications services by voice, with the goal of eliminating the need for a customer to learn the different access codes for existing or new features. In addition, the service provided voice confirmation that the service requested was being turned on or off.

The procedure for using VIP is for the customer to dial an abbreviated access code (e.g., 3 digits), an then hear a prompt of the form:

> "Please say, the name of the feature you want, or say 'HELP' for a list of the services you subscribe to, now."

The user then speaks the name of the service and receives confirmation of connection to that service. The services available through VIP, and the associated voice commands are as follows:

Service	Voice Command
Call Forwarding	Call Forwarding
Continuous Redial	Redial
Last Call Return	Return Call
Call Rejection	Call Rejection
Caller ID Blocking	Block ID
Access to Messaging Services	Messages
Temporary Deactivation of Call Waiting	Cancel Call Waiting

Based on a series of customer trials, the following results were obtained:

1) 84% of the users preferred VIP over their present method.
2) 96% of the users were comfortable with the idea of speaking to a machine.
3) Most users felt that the primary benefit of VIP was not having to remember multiple codes or procedures.
4) 75% of users tried different services with VIP more often, or were willing to try services they had never tried before.

In addition to the applications to telecommunications, there are several interesting speech recognition applications in the other segments of the market. One such application is a voice repertory dialer which acts as a hands-free, eyes-free, adjunct for a cellular phone, and provides the capability of calling a preassigned number via a voice alias, e.g., "call home," or "call the office." An arbitrary number could also be voice dialed but this requires a full connected digit recognition capability, which generally has not been available within cellular recognizers.

E. Ease of Use Issues

A key aspect in the success of voice recognition applications is how well the human–machine interface has been designed so that the recognition system is truly easy to use. The goal of the human factors design is to delight the customer with the ease of use and the apparent simplicity of the task. The human factor enters through the judicious design and use of prompts and reprompts, as well as in the mode (auditory, visual, tactile), timing, and content of feedback to the user. By way of example, consider the canonic prompt for a voice service for the XYZ company:

Ideal Prompt: "Welcome to XYZ service. How may I help you?

Ultimately this ideal prompt will become reality; at the current time it would give the user far too much opportunity to say things that the recognizer cannot handle. Hence a more realistic prompt might be:

Realistic Prompt: "Welcome to XYZ service. Please say A, B, C, or D now."

Experience has shown that an improved prompt would be:

Improved Prompt: "Welcome to XYZ service. What type of service would you like to access? (Pause) Please say A, B, C, or D now."

The benefits of the improved prompt are the following:

1) Experienced users will barge-in at the pause, thereby improving system throughput.
2) Novice users listen to the entire message, getting a better picture of the way in which the transaction is carried out.
3) All users find the question helpful in understanding the transaction with the result that there are fewer "no response" cases.
4) Customers complete the transaction faster.
5) The recognition system is rated higher by customers.

For reprompting, either after a failed recognition, or in response to improper input, a standard reprompt command might be:

Standard Reprompt: "Your response was not understood. Please say ... now."

An improved reprompt is:

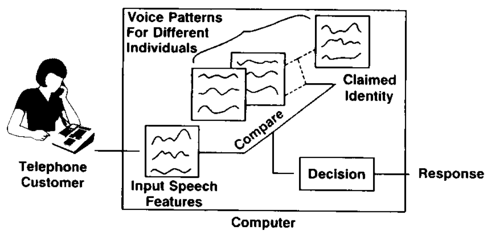


Fig. 39. Block diagram of a speaker verification system.

Improved Reprompt: "Sorry, please repeat. (Pause) Please say ... now."

The benefits of the improved reprompt include:

1) much faster performance,
2) better customer acceptance,
3) more barge-in (both early, and at the pause),
4) better conversational style and pace,
5) less customer frustration with repeats.

Good human factor leads to getting the best performance from a speech recognition application.

VI. Speaker Verification [27]–[29]

The basic problem of speaker verification is to decide whether or not an unknown speech sample was spoken by the individual whose identity was claimed. The problem is similar to that of speech recognition in which the problem is to normalize out, in some sense, the individual speaker and extract the message content of the speech. Here, the problem is to normalize out, in some sense, the message content and extract information about the individual speaker. Because of the similarities of these two problems, the processing for speaker verification is similar (with some small differences) to that of speech recognition.

Figure 39 shows a canonic speaker verification system. The customer, wishing to be verified, provides a claimed identity (which enables the system to retrieve the voice pattern corresponding to the identity claim), and a voice sample. The speech features of the customer's sample are compared, using a time-alignment procedure similar to the one used for speech recognition, to the voice pattern corresponding to the claimed identity, and, if a suitable match is obtained, the identity claim is verified.

A. Generic Applications of Speaker Verification

The major area of application for speaker verification is in access control to information, credit, banking, machines, computer networks, private branch exchanges (PBX's), and even premises. Thus the concept of a "voice lock" that prevents access until the appropriate speech by the authorized individual(s) is "heard" is made a reality by speaker-verification technology.

B. Speaker Verification Technology [27]

Figure 40 shows a block diagram of an integrated speaker verification system in which the customer wishing to be verified provides a claimed identity (in order to access the appropriate stored voice pattern), the spoken phrase suitable to the verification system, and the transaction requested. A comparison of the spoken phrases (suitably time-aligned) with the appropriate stored voice pattern provides a comparison score. Depending on the transaction requested, the decision to accept or reject the identity claim is made and sent back to the customer via a computer speech answer-back system. Thus for banking transactions, a much lower degree of match would be required to check an account balance than would be required to withdraw funds.

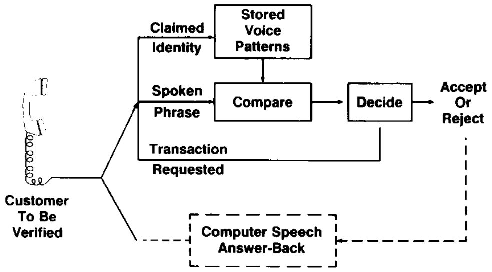


Fig. 40. Block digram of an integrated speaker verification system using computer speech answer-back to provide user feedback.

A speaker verification system can make two types of errors; it can reject a true customer (Type I error) or it can accept an imposter (Type II error). The goal of most verification systems is to try to bound Type I errors (e.g., <0.5%) while minimizing Type II errors (e.g., at 10%). Often, in laboratory testing, performance scores are given for equal rates of Type I and Type II errors.

The performance of a speaker verification system is a very sensitive function of several factors, including:

1) the microphone used in the recording, especially when both carbon button and electret microphones are intermixed in training and testing;
2) the transmission channel;
3) the background noise;
4) the speaker condition (e.g., colds);
5) the usage condition, i.e., speakerphone, cordless phone, cellular phone.

Various technical solutions have been proposed for these problems; however, there is no perfect solution which provides a robust system with high performance in the field.

C. Performance of Speaker Verification Systems

Table 3 provides a summary of the performance achieved in laboratory evaluations of a speaker verification system. The particular system that gave the results shown in Table 3 used digit sequences for both training and testing (in particular, 7-digit test utterances were used). The performance scores shown in the table are equal error rate scores (in percent), which means that these are the scores for a decision threshold set (experimentally, based on measured errors for each talker) so that the true customer is rejected the same percentage of the time an average imposter gets

Table 3 Performance of a Speaker verification System Using Connected Digit Sequences as Input Strings

Adaptation	Text	
	Independent	Dependent
Without	3.0%	0.8%
With	2.2%	0.3%

accepted. Shown are results for both *text-independent trials* (those in which the customer can speak any arbitrary 7-digit sequence and the machine does not know the digits), and for *text-dependent trials* in which the machine instructs the customer as to the exact digit sequence to speak. Also shown are results without and with adaptation, over time, of the reference pattern to the changes in speaking patterns of the user. Clearly, the performance in the text-dependent mode is far superior to that of the text-independent mode because the machine can exploit the known dynamics and spectral content of the speech utterance precisely in making its decision.

Also shown in the table are results without and with *adaptation* to the changing talker characteristics (individual speaker voice patterns) over time. The best performance of 0.3% equal error rate is achieved with text-dependent mode and with adaptation; the loss in performance without adaptation (to 0.8% equal error rate) is significantly smaller than the loss in performance in the text-independent mode (to 2.2% equal error rate), or to 3.0% equal error rate when adaptation is not used. Hence, the extra information obtained from foreknowledge of the spoken digit string is significantly greater than that obtained from adaptation to the changes in the talker's speech patterns.

D. Telecommunications Applications of Speaker Verification

Although the technology for speaker verification has been around, and well understood, for a number of years, there has been essentially no commercialization of the technology until recently. This is because security is a feature that most customers are unwilling to pay for—until a break-in occurs. With the opening up of computers, networks, and other telecommunications systems, the need for security has grown to the point where speaker verification is now an attractive alternative to electronic security for:

1) ATM (Automated Teller Machines), using smart cards to store voice patterns using on the order of 20 000 bits of storage, as announced by NCR.
2) PBX Services, to provide protection against improper use of PBX for calls made from outside of the office environment.
3) Network services, where speaker verification provides access to a range of telecommunications services such as name dialing using voice aliases, teletravel information, FAX services, etc.
4) Computer systems, as an adjunct to electronic security as provided by passwords.

It is anticipated that speaker verification will appear in applications in each of the above areas over the next few years.

VII. Spoken Language Translation [30]

An obvious extension to research in speech synthesis and speech recognition, is spoken language translation which holds the promise of providing the capability of having a dialog between 2 (or more) speakers neither of whom speak the same (or a common) language.

Programs in language translation go back as early as the 1960's when the US Government funded a program for natural language translation of text. The results of this program were mixed, at best, with poor quality of translation and with no clear path to success emerging. The next major milestone began in the 1980's when NEC and ATR in Japan began long-range programs on so-called interpreting telephony, i.e., speech-based language translation. The key to these efforts was an understanding of the role of syntax and semantics in both the recognition and language translation algorithms, leading to the creation of low-perplexity systems which achieved reasonably good success on highly limited tasks. In the 1990's, AT&T began a program in spoken language translation (in conjunction with Telefonica of Spain), leading to the VEST (Voice English–Spanish Translator) system, a limited task domain, medium-perplexity system.

Spoken language translation is an ongoing area of research throughout the world. There are four requirements which are necessary (and often sufficient) for a spoken language translation program to succeed, and these are:

1) a limited task domain: The task domain for the VEST system is banking and currency exchange. This feature is necessary to keep the task perplexity low and to provide the capability of high accuracy in the speech recognition part of the program.
2) a common language model for both recognition and parsing of text: This requirement is necessary in order to create an accurate model of language, and to keep the language model for translation in synchrony with the language model for recognition.
3) expertise in speech recognition and synthesis: This requirement is necessary to make the spoken language translation problem viable.
4) high speed processing capability: This requirement is necessary to insure real-time transactions so that the system can actually be used, tuned, and developed.

Based on these requirements, AT&T and Telefonica jointly built the VEST system, which is described in more detail in the next section.

A. Spoken Language Translation Technology [30]

Figure 41 shows a block diagram of a spoken language translation system, as used in the VEST project (where the two languages are English and Spanish). The system consists of:

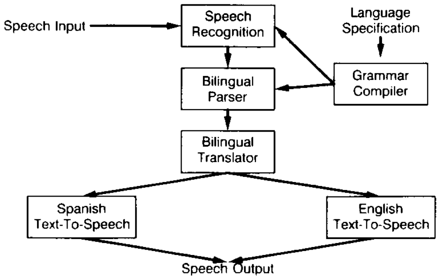


Fig. 41. Flow digram of the VEST (Voice English–Spanish Translator) system.

1) A dual language speech recognizer which automatically determines both the spoken language, and the best sentence in that language. The recognizer is controlled by a grammar network, generated from a task language specification, which is common to both the recognizer and a bilingual parser.
2) A bilingual parser which parses the English or Spanish text according to the language model for the task.
3) A bilingual translator which maps the parse tree in one language to a parse tree in the companion language according to a set of language translation rules appropriate to the task and the parse tree.
4) A pair of TTS modules, one for Spanish one for English, which speaks out the translated version of the spoken input request in the companion language.

The VEST system runs on an array DSP processor with 128 floating-point DSP's (AT&T DSP 32C) performing all the computation for recognition and language translation, and with a workstation (a SPARC II) doing the TTS for both English and Spanish. The common language model, used for recognition and language translation, was created using a standard grammar compiler, available at Bell Labs. The system has a vocabulary of 453 words, combined; uses 401 production rules (hand written) to generate the language model; uses 1228 rules in the context free grammar for parsing; has a 1600 state, 6400 node, 7500 transition finite state grammar for recognition, with a perplexity of 86; and uses 877 translation rules (hand written) to translate between languages. The entire system was informally evaluated and achieved 96% semantic sentence accuracy for a single talker. (The system is speaker-trained.) The VEST system was successfully demonstrated at the 1992 World's Fair in Seville, Spain, by Telefonica, for a 6 month period with about 6 sets of trained operators.

B. Issues in Spoken Language Translation

Although a modest degree of success has been achieved in building spoken language translation systems, there remains a great deal to be done before such systems can be applied to real problems in telecommunications. Among the issues which must be investigated and resolved are the following:

1) the need for common tasks with increasing levels of difficulty so as to evaluate progress and problems which need to be solved;
2) the need for large amounts of training and testing data to train the recognition systems, and to evaluate performance of each part of the system;
3) the need for a formal evaluation methodology so as to separate overall system performance scores from scores for each component of the system: we also need to understand how to maintain a user dialog in the face of small and gross errors in either recognition or translation.
4) the need for lots of computing capability to keep up with recognition and translation in real-time, an absolute necessity for a system for spoken language translation;
5) the need for more statistical and knowledge-based paradigms to automate the processes of language model generation, and language translation rule generation.

VIII. Acoustic Challenges in Telecommunications [31]–[33]

One of the challenges of universal communications is to provide high-quality voice communications in different environments (e.g., restaurants, train stations, airports, offices, homes), in different mobility modes (e.g., walking, jogging, driving, stationary), and in different group situations (e.g., one-on-one, one-on-many, many-on-many). There are three acoustic challenges that must be met to provide this universal voice communications capability, including:

1) advanced microphone technology which has the capability of finding and localizing a sound source (i.e., talker) as well as suppressing noise sources in the background [32];
2) advanced loudspeaker technology which has the capability of focusing sound in a desired field (e.g., around a single listener, or in a cone of listeners);
3) advanced echo cancellation which has the capability of eliminating feedback paths from the loudspeaker to the microphone [31], [33].

In the area of microphone technology a wide array of designs have been proposed and tested, including:

1) Differential microphones which provide on the order of up to 20-dB cancellation of a far end noise source while enhancing a signal source in the near field of the microphone.
2) Adaptive microphones which are capable of locking onto and tracking a signal source and keeping approximately constant volume from the source.
3) Array microphones which automatically, and adaptively, configure themselves to find and track signal sources while rejecting undesired noise and reverberation.

Current research is focused on the creation of "smart" microphones which integrate the capabilities of differential, adaptive, and array microphones in a single configuration which automatically and adaptively configures itself to find and track signal sources, at the same time finding, tracking, and canceling noise sources. Similarly, we can envision "smart" loudspeakers which automatically and adaptively configure themselves to focus sound in a specified region of space.

Taken together, a system with smart microphones, smart loudspeakers, and appropriate echo cancelers, holds the promise of providing echo-free, noise-free, high-quality (broad-bandwidth) speech communications between groups of people, anywhere, and at anytime.

A. Vehicle-Focused Speech Processing

By way of example, consider the capabilities of controlling and communicating in a vehicular environment, that are enabled by voice processing and associated electroacoustic transducers. One set of possibilities is shown in Fig. 42. It can be seen that the opportunities include:

1) speech-based alerting, based on coding or TTS, of problems associated with car operation or the audio system within the car;
2) speech control, via voice recognition, of various aspects of instrumentation including the radio, the comfort features of the car, lights, wipers, speed control, and turn signals;
3) voice lock access to the car, via speaker verification, including restricted access to the trunk and glove box: based on verification, a personalized set of driving options, such as seat and mirror positions, could be created for each driver of the car;
4) use of cellular telephony for hands-free, eyes-free, voice telephony within the car, including repertory dialing of commonly called numbers by voice commands;
5) use of hands-free microphone for telephony and voice control of car features;
6) use of acoustic conditioning to minimize car noise and to create a good acoustic environment for listening to audio broadcasts in the car.

It is clear that the car of the future could look and feel significantly different than the car of today, based on voice processing technology.

IX. Summary

As we head inexorably toward the turn of the century, we see rapid progress toward the vision of universal communications, especially in the area of voice processing. As mentioned several times in this paper, the revolution in VLSI, especially in the area of DSP chip technology, has fueled many of the advances by providing the computational power to perform a wide range of voice and image processing operations in real time on a single chip. This point is illustrated in Fig. 43 which shows a plot of the single DSP chip CPU instructions per second versus time (from 1980 to 1995), along with individual points showing the CPU capability needed for various voice and image processing tasks. It can be seen that in 1993, all but the most advanced large-vocabulary speech recognition and the HDTV receiver tasks can be implemented on a single DSP chip. Thus the implication of this is that with program downloading capability, a single DSP chip could be multiplexed to perform a wide range of voice processing, image processing, and communications functionality, thereby providing a strong platform for the personal communicator of the year 2001.

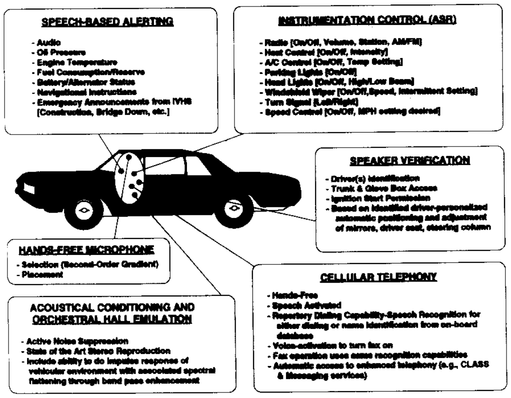

SPEECH-BASED ALERTING
- Audio
- Oil Pressure
- Engine Temperature
- Fuel Consumption/Reserve
- Battery/Alternator Status
- Navigational Instructions
- Emergency Announcements from IVHS [Construction, Bridge Down, etc.]
INSTRUMENTATION CONTROL (ASR)
- Radio [On/Off, Volume, Station, AM/FM]
- Heat Control [On/Off, Intensity]
- A/C Control [On/Off, Temp Setting]
- Parking Lights [On/Off]
- Head Lights [On/Off, High/Low Beam]
- Windshield Wiper [On/Off,Speed, Intermittent Setting]
- Turn Signal [Left/Right]
- Speed Control [On/Off, MPH setting desired]
SPEAKER VERIFICATION
- Driver(s) Identification
- Trunk & Glove Box Access
- Ignition Start Permission
- Based on Identified driver-personalized automatic positioning and adjustment of mirrors, driver seat, steering column
HANDS-FREE MICROPHONE
- Selection (Second-Order Gradient)
- Placement
ACOUSTICAL CONDITIONING AND ORCHESTRAL HALL EMULATION
- Active Noise Suppression
- State of the Art Stereo Reproduction
- Include ability to do impulse response of vehicular environment with associated spectral flattening through band pass enhancement
CELLULAR TELEPHONY
- Hands-Free
- Speech Activated
- Repertory Dialing Capability-Speech Recognition for either dialing or name identification from on-board database
- Voice-activation to turn fax on
- Fax operation uses same recognition capabilities
- Automatic access to enhanced telephony (e.g., CLASS & Messaging services)

Fig. 42. Examples of how voice processing and electroacoustic design could be incorporated into an automobile to improve the performance and make the features of the automobile easier to use.

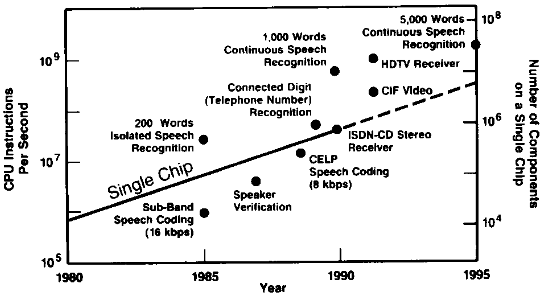


Fig. 43. Plot of the signal processing requirements for several speech and image processing applications as compared to the capability of single-chip DSP processors from 1980 to 1995.

Acknowledgment

The technology and applications described in this paper are the results of the research and development efforts of a large number of individuals within Bell Labs and AT&T, as well as outside of AT&T. Although the number of individuals who have contributed to the technology is too large to mention here, several individuals must be acknowledged for their personal contributions to the material presented here. These include J. Flanagan for his work in defining the personal communications technology; N. Jayant, B. Atal, R. Cox, and P. Kroon for their work in speech coding; J. Olive and J. Bachenko for their work in speech synthesis; J. Wilpon, B. Perdue, D. Roe, and D. Thomson for their work in speech recognition; A. Rosenberg for his work in speaker verification; and D.

Berkley, M. Sondhi, G. Elko, and J. West for their work in electroacoustics. Many of the ideas, figures, data, and much inspiration for the material contained in this paper, are due to these colleagues.

References

[1] B. S. Atal, "Speech processing based on linear prediction", in *Encyclopedia of Physical Science and Technology*, vol. 13. New York: Academic Press, 1987, pp. 219–230.

[2] J. H. Chen, R. V. Cox, Y. C. Lin, N. S. Jayant, and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Select. Areas Commun.*, vol. 10, no. 2, pp. 830–849, June 1992.

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

[4] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[5] P. Kroon and B. S. Atal, "Strategies for improving the performance of CELP coders at low bit rates," in *Proc. IEEE Int. Conf.on Acoustics, Speech, and Signal Processing*, Apr. 1988, pp. 151–154.

[6] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, 1979.

[7] Y. Shoham, "Low-rate speech coding based on time-frequency interpolation," in *Proc. Int. Conf. on Spoken Language Proc.* (ICLSP'92), Oct. 1992, pp. 37–49.

[8] J. Allen, "Synthesis of speech from unrestricted text," *Proc. IEEE*, vol. 64, pp. 422–433, 1976.

[9] J. Allen, S. Hunnicutt, and D. H. Klatt, *From Text to Speech: The MIT Talk System*. Cambridge, England: Cambridge Univ. Press, 1987.

[10] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, pp. 452–459, 1976.

[11] J. L. Flanagan, "Computers that talk and listen: Man–machine communication by voice," *Proc. IEEE*, vol. 64, pp. 405–415, 1976.

[12] J. L. Flanagan and L. R. Rabiner, *Speech Synthesis*. Stroudsberg, PA: Dowden, Hutchinson and Ross, 1973.

[13] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, pp. 127–143, 1964.

[14] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, no. 3, pp. 737–793, Sept. 1987.

[15] J. P. Olive and M. Y. Liberman, "A set of concatenative units for speech synthesis," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, J. J. Wolf and D. H. Klatt, Eds. New York: Ameri. Inst. of Physics, 1979, pp. 515–518.

[16] M. H. O'Malley, D. K. Larkin, and E. W. Peters, "Beyond the reading machine: What the next generation of intelligent text-to-speech systems should do for the user," in *Proc. Speech Tech. '86*, 1986, pp. 216–219.

[17] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene, "Perception of synthetic speech generated by rule," *Proc. IEEE*, vol. 73, pp. 1665–1676, 1985.

[18] L. R. Bahl, F. Jelinek, and R. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

[19] K. F. Lee, *Automatic Speech Recognition, the Development of the SPHINX System*. Boston, MA: Kluwer, 1989.

[20] H. Ney, "The use of a one stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 263–271, 1984.

[21] L. R. Rabiner, "A tutorial on hidden markov models and its applications to speech recognition," *Proc. IEEE*, vol. 72, no. 2, pp. 257–286, 1989.

[22] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[23] L. R. Rabiner and S. E. Levinson, "Isolated and connected word recognition—Theory and selected applications," *IEEE Trans. Commun.*, vol. COM-29, no. 5, pp. 621–659, 1981.

[24] D. R. Reddy, "Speech recognition machine: A review," *Proc. IEEE*, vol. 64, pp. 502–531, 1976.

[25] A. Weibel and K. F. Lee, Eds., *Readings in Speech Recognition*. San Mateo, CA: Morgan Kaufman, 1990.

[26] V. W. Zue, "The use of speech knowledge in automatic speech recognition," *Proc. IEEE*, vol. 73, no. 11, pp. 1602–1615, 1985.

[27] A. E. Rosenberg, "Automatic speaker verification: A review", *Proc. IEEE*, vol. 64, pp. 475–487, 1976.

[28] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Computer, Speech, and Language*, vol. 22, pp. 143–157, 1987.

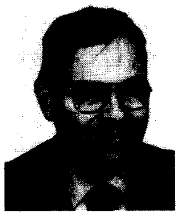
[29] F. K. Soong, A. E. Rosenberg, B. H. Juang, and L. R. Rabiner, "A vector quantization approach to speaker recognition," *AT&T Tech. J.*, vol. 66, pp. 14–26, 1987.

[30] D. B. Roe, P. J. Moreno, R. W. Sproat, F. C. Pereira, M. D. Riley, and A. Macarron, "A spoken language translator for restricted-domain context-free languages," *Speech Commun.*, vol. 11, pp. 311–319, 1992.

[31] S. L. Gay, "Fast converging subband acoustic echo cancellation using RAP on the WE DSP 16A," in *Proc. ICASSP'90* (Albuquerque, NM, Apr. 1990), pp. 1141–1144.

[32] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, and M. M. Sondhi, "Autodirective microphone systems," *Acoustica*, vol. 73, pp. 58–71, Feb. 1991.

[33] M. M. Sondhi and W. Kellermann, "Adaptive echo cancellation for speech signals," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992.



Lawrence R. Rabiner (Fellow, IEEE) was born in Brooklyn, NY, on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in June 1964, and the Ph.D. degree in electrical engineering in June 1967, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964, he participated in the cooperative plan in electrical engineering at AT&T Bell Laboratories, Whippany and Murray Hill, NJ. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is Director of the Information Principles Research Laboratory at AT&T Bell Laboratories, Murray Hill, and is engaged in research on speech communications and digital signal processing techniques. He is coauthor of the books *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), *Multirate Digital Signal Processing* (Prentice-Hall, 1983), and *Fundamentals of Speech Recognition* (Prentice-Hall, 1993).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, the National Academy of Engineering, the National Academy of Sciences, and a Fellow of the Acoustical Society of America, and AT&T Bell Laboratories. He is a former President of the IEEE Acoustics, Speech and Signal Processing Society, a former editor of the IEEE TRANSACTIONS ON ACCOUSTICS, SPEECH, AND SIGNAL PROCESSING and a former member of the IEEE PROCEEDINGS Editorial Board.