2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94) Kyoto Research Park, Kyoto, Japan September 26 - 27, 1994

THE ROLE OF VOICE PROCESSING IN TELECOMMUNICATIONS

Lawrence R. Rabiner

AT&T Bell Laboratories
600 Mountain Avenue, Murray Hill, New Jersey 07974
USA

ABSTRACT

During the decade of the 1990's, the fields of communications, computing, and networking are coming together in the form of personal information/communication terminals, and in the associated services (so-called Personal Communications Services — PCS). Several technologies will play major roles in this communications revolution, but one of the key ones will be voice processing. In this paper we review several voice processing technologies, discuss current capabilities and the associated applications, and try to forecast where we see progress being achieved in the next decade and what applications will become commonplace as a result of the increased capabilities. We show how progress in voice processing is accompanied and stimulated by progress in microelectronics (memory and processing power of single chip architectures), and how, by the 21st century, telecommunications will have made major advances as a result of the use of voice processing.

1. INTRODUCTION

The vision of telecommunications in the twenty-first century is to provide seamless, high-quality communications, between people (or groups of people) and machines, anywhere, anytime, and at a reasonable price. In order to understand this vision, we must first think about the implications of each piece of this vision statement. The elements of the vision are as follows:

Communications

The idea of communications is any type of information exchange between people and/or machines. Hence communications includes voice calls, voice messages, video calls, video messages, electronic mail, FAX, handwritten notes (sent electronically), and data files (e.g., spreadsheets, word processing documents, charts, databases, etc.).

Seamless Communications

The idea behind seamless communications is that the user can integrate any form of communication with any other form of communication in what appears (to both the user and the person(s) with whom he or she is communicating) to be an effortless process and which does not degrade the quality of the communications. For example, in the middle of a voice call, the caller should be able to scribble a note on an electronic message pad and send it off so that it is received simultaneously by the other party, without in any way affecting the quality or the protocol of the voice call. Seamless communications also implies the capability of converting one form of communications to another, more convenient form, as appropriate. In this manner, an e-mail or FAX message could be converted to a voice message when the receiving party does not have convenient access to a display.

High Quality Communications

The idea behind high-quality communications is that the user does not perceive any unnatural degradation in the signal, no matter what the speaking environment or transmission medium. Thus, in theory, a voice call should sound the same over a wired or wireless network, and a video call should look and sound the same over both networks. In reality, most users expect some degradation when communicating in noisy environments (e.g., airplanes, train terminals) and therefore high-quality communications is defined relative to user expectations and perceptions. High quality communications also implies ease-of-use and convenience of the communications systems, e.g., control of communication flow via voice commands rather than touch-tone pads, etc.

Anywhere Communications

The idea behind anywhere communications is that a user can access the same communications environment and resources anywhere in the world, thereby providing the capability of staying in touch no matter where the user happens to be. The implications of anywhere communication include the existence of a worldwide wireless infrastructure which couples directly into a worldwide long distance network so that a personal communicator could be used both indoors (wired or wireless) or outside (wireless). Anywhere communications also implies a roving capability (since the person you want to communicate with could be anywhere) along with the availability of an up-to-date network database of the current location of every potential user of the system.

Anytime Communications

The idea behind anytime communications is that a user can communicate when he or she wants to—even if the person to whom the communications is being sent is unavailable (or doesn't want to accept communications at the time). The key concept for anytime communications is integrated messaging, whereby every user has a single access number

for all communications with an integrated attached mail box. Such a system would have to be able to determine the nature of all incoming communications and direct it appropriately. It would also have to be capable of seamlessly converting voice and video calls to voice and video messages when the recipient is unavailable or doesn't want to receive calls directly.

Communications at a Reasonable Price

The key idea behind communications at a reasonable price is that all of the software and processing for the communications fit on a low cost, low power processor, so that a wireless personal information terminal could be used a full day on a single set of batteries without the need to recharge or change batteries during the day. Low cost communications also implies the existence of a low cost, high data rate, infrastructure within the network (the so-called information superhighway) so that network services can be provided at a reasonable price.

To realize this communications vision there has to be a steady stream of progress in a number of key areas including computer science, microelectronics, batteries, networking, and communications. In this paper we restrict our attention to one key communication technology, namely voice processing. For each area of voice processing we review our current research capability, discuss the telecommunication applications that are currently available, and point out some of the challenges in realizing the dream of universal voice communications.

2. AREAS OF VOICE PROCESSING [1]

Voice processing can be conveniently segmented into the following areas:

speech coding – compressing the information in a speech signal for efficient transmission or storage

speech synthesis – conversion of arbitrary ASCII text to speech in order to transmit a message from a machine to a person

speech recognition – extraction of the message information from spoken inputs in order to control the actions of a machine in response to spoken commands

speaker recognition – verification of the claimed identity of a speaker so as to restrict access to information, networks, or physical premises.

In each of these areas we discuss current capabilities, show typical telecommunications applications, and give our views of the challenges for the future.

2.1. Speech Coding [2]

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.

2.1.1. Coding Technology

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. Figure 1 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. Several variants on the basic speech coder model of Figure 1 have been studied including the vocoder model, the multipulse model, and the stochastic (CELP) model.

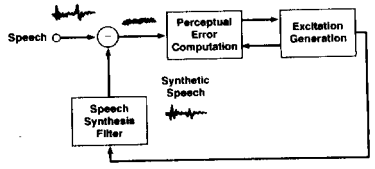

Figure 1: Block diagram of a perceptually driven speech analysis/synthesis system.

2.1.2. Coder Evaluation

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 transmis-

sion 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 msec) 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.

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, Figure 2 shows plots of speech quality (as measured in terms of mean opinion scores (MOS)) for a range of coders spanning bit rates from 64 Kbps down to 2.4 Kbps. (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 Kbps
2) adaptive differential pulse code modulation (ADPCM) at 32 Kbps
3) low delay code-excited linear prediction (LD-CELP) at 16 Kbps
4) vector sum excitation linear prediction (VSELP) at 8 Kbps (more precisely 7.950 Kbps)
5) code excited linear prediction (CELP) at 4.8 Kbps
6) linear predictive coding (LPC10 E) at 2.4 Kbps

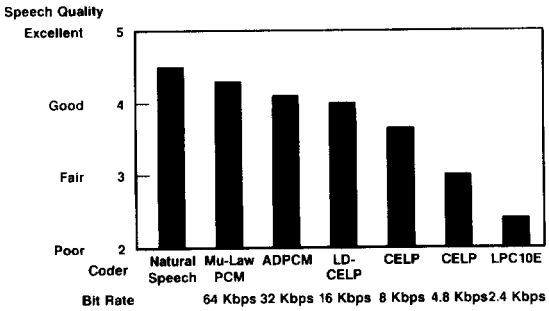

Figure 2: Speech quality mean opinion scores of several coders as a function of bit rate.

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 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 Figure 2, 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 Kbps to 64 Kbps 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 Kbps, 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 Kbps, 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.

2.1.3. Telecommunication Applications of Telephone Bandwidth Coders

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.

We now examine each of these areas.

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.

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.).

3) 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 Mbit), and with appropriate low rate speech coders (e.g., 6.6 to 13.0 Kbps range) on the order of 5 minutes of coded speech can be stored on a single 4 Mbit chip, and on the order of 20 minutes of coded speech can be stored on a 16 Mbit chip. Hence the usual 30 minute 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.

4) 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 Kbps 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 Kbps (the so-called Secure Telephone Unit (STU III) device), and the second (called the Secure Telephone Device 3600) running just at a 4.8 Kbps rate. The STU-III uses LPC10 E coding at 2.4 Kbps and CELP coding at 4.8 Kbps. The STD 3600 uses RCELP coding at 4.8 Kbps.

2.1.4. 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 Kbps range. Current research is focusing on lowering the bit rate to 16 Kbps while maintaining high-quality so as to provide audio/video teleconferencing at 128 Kbps with 112 Kbps provided for video coding, and 16 Kbps for high-quality audio coding.

2.1.5. Challenges in Voice Coding

There are essentially three challenges in voice coding to achieve the communications vision of the twenty-first century. These are the following:

- achieve high-quality coding at cellular and lower rates, i.e., MOS scores of 4.0 or higher for data rates of 8 Kbps down to 2.4 Kbps with reasonable delay and reasonable complexity
- provide the capability of speeding up or slowing down coded messages without seriously affecting the perceived message quality
- achieve robust coding performance in noisy and reverberant environments.

Current research is aimed at solving each of these problems.

2.2. Speech Synthesis [3, 4]

The goal of speech synthesis is to provide a broad range of capability for having a machine speak information (respond) to a user over an appropriate communications channel.

2.2.1. Speech Synthesis Technology

Speech synthesis systems can be realized as either simple concatenation systems, as shown in Figure 3, or as full TTS systems, as shown in Figure 4. 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.

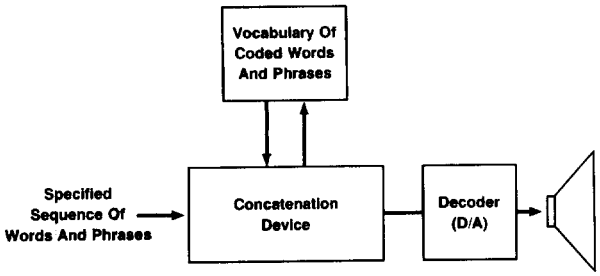

Figure 3: Block diagram of simple voice response system (D/A: digital-to-analog).

For the full TTS system of Figure 4, 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 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. 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.

2.2.2. Telecommunication Applications of Speech Synthesis

A number of interesting and important applications of TTS have evolved and are currently in use. These include the following:

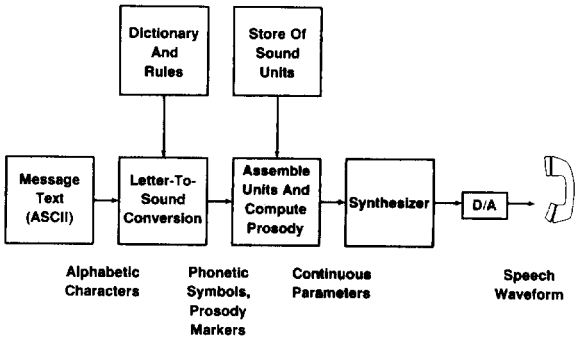

Figure 4: Block diagram of full text-to-speech synthesis system.

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) 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.

4) Business locater 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.

5) 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.

6) Banking services providing the customer with access to and control of bank accounts including account status, check status, and bill paying options.

2.2.3. Challenges in Speech Synthesis

Our current capability in TTS is a complete system for English which is capable of taking an arbitrary ASCII script and producing highly intelligible (synthetic quality) speech in English (with excellent intelligibility of proper names and addresses) for a male speaker, along with passable quality

for a female voice. 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 non-plane 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.

2.3. Speech Recognition

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.

2.3.1. Speech Recognition Technology [5]–[8]

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 Figure 5. 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 system of Figure 5 can assume many forms, e.g., a template based approach or a sophisticated statistical model (the hidden Markov model), and can be applied to a broad range of problems including isolated word/phrase recognition, connected word recognition, and even continuous speech recognition. Although the underlying techniques are often quite sophisticated, the basic pattern recognition model is the basis for almost all currently used methods.

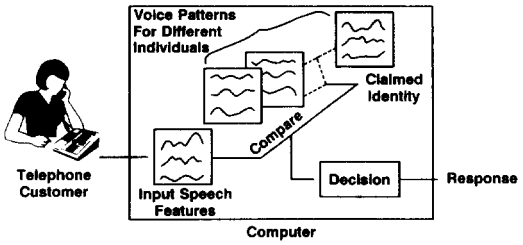

Figure 5: Simple block diagram of pattern-recognition model for word recognition.

2.3.2. Telecommunications Applications of Speech Recognition

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, information that was previously not possible, the customers are generally cooperative, and quite forgiving of technology limitations.

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 O+ 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).

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.

2.3.3. Challenges for Speech Recognition

Our current capability for recognizing speech is highly dependent on a variety of factors. For isolated words and phrases, recognition systems have achieved error rates below 5% (often well below this rate); however such systems are often fragile and their performance degrades significantly in noisy backgrounds, when different transducers are used, etc. For connected word recognition of fluent speech, systems have achieved error rates as low as 0.3% for digit sequences and 3% for a 1000 word vocabulary and a highly constrained task.

The major challenge in speech recognition is to learn how to build speech recognition (or more broadly speech understanding) systems which are easy-to-use, robust to environment, channel, transducer, and talkers, and which work in virtually any language, for virtually unlimited vocabularies, and with unconstrained syntax. This challenge is long term and fairly open-ended, and probably will not be solved for several decades. In the next several years we can define more realistic challenges which will enable a wide range of telecommunication applications of speech recognition. These include the following:

- connected digit recognition that is highly accurate (digit error rate <0.5%) and is robust to the talker, background noise, telephone handset and transmission channel
- high performance (>95% accuracy per transaction) speech recognition based on subword speech units for medium size vocabularies (100–2000 words) for simple tasks (stock price quotations, directory listing retrieval, etc.)
- robust rejection of extraneous speech (i.e., speech with no valid commands) and background sounds.

2.4. Speaker Verification [9]

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.

2.4.1. Speaker Verification Technology

Figure 6 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.

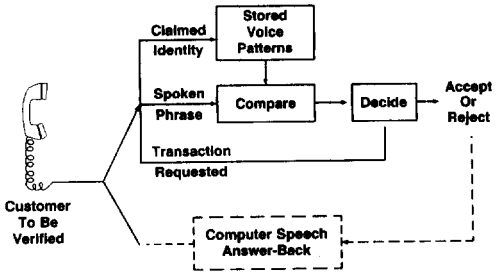

Figure 6: Block diagram of an integrated speaker verification system using computer speech answer-back to provide user feedback.

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.

2.4.2. 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.

2.4.3. Speaker Verification Challenges

At the current time speaker verification by voice can be very accurate (with equal error rates reported as low as 0.3%) under controlled laboratory conditions (i.e., a cooperative user speaking a machine specified string of digits, in a controlled noise background, with a fixed transducer, and a fixed transmission channel), with adaptation to the talkers speech over time. However it remains a formidable challenge to build a "real world" speaker verification system that is highly accurate for accepting true speakers and rejecting imposters, and one that is insensitive to the variability inherent in real-world situations. A second key challenge is to develop a verification system that can be used with minimal training and that can effectively adapt, over time, to the talker.

3. THE COMPUTING AND NETWORKING ENVIRONMENT OF THE FUTURE

One of the major factors spurring the communications revolution is the inexorable rate of increase in memory capacity and processor capability of single chip components, as well as the rapid rate of increase of the networking capability in almost every facet of the workplace. It is predicted that by the year 2001 memory chips will have a 1 Gbit capacity, dsp chips will be capable of processing upwards of 4000 MIPS, and even single chip microprocessors (of the Pentium or Power PC line) will be able to provide 1000 MIPS or higher processing speeds.

In the area of networking the desktop in the office, as well as the desktop in the home, will have fiber optic links carrying information and data at rates upwards of 1 Gbit/second, i.e., fast enough to send all the text in a large encyclopedia in less than a second.

The effect of this enormous increase in memory, computing, and networking power is that speech processing will eventually become a low cost feature of most integrated communications systems of the future. Hence there will be essentially no real computational limitations in our ability to devise new signal processing algorithms for speech processing, or to implement them in low cost silicon.

REFERENCES

[1] L. R. Rabiner. Applications of voice processing to telecommunications. *Proc. IEEE*, 82(2):199–228, February 1994.

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

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

[4] D. H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Amer.*, 82(3):737–793, Sept. 1987.

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

[6] D. R. Reddy. Speech recognition by machine: A review. *Proc. IEEE*, 64:502–531, 1976.

[7] A. Weibel and K. F. Lees, editors. *Readings in Speech Recognition.* Morgan Kaufman, San Mateo, CA, 1990.

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

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