Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9991–9998, October 1995
Colloquium Paper

This paper is a condensed version of one that was presented at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Voice-processing technologies—Their application in telecommunications

JAY G. WILPON

AT&T Bell Laboratories, Murray Hill, NJ 07974

ABSTRACT **As the telecommunications industry evolves over the next decade to provide the products and services that people will desire, several key technologies will become commonplace. Two of these, automatic speech recognition and text-to-speech synthesis, will provide users with more freedom on when, where, and how they access information. While these technologies are currently in their infancy, their capabilities are rapidly increasing and their deployment in today's telephone network is expanding. The economic impact of just one application, the automation of operator services, is well over $100 million per year. Yet there still are many technical challenges that must be resolved before these technologies can be deployed ubiquitously in products and services throughout the worldwide telephone network. These challenges include: (*i*) *High level of accuracy.* The technology must be perceived by the user as highly accurate, robust, and reliable. (*ii*) *Easy to use.* Speech is only one of several possible input/output modalities for conveying information between a human and a machine, much like a computer terminal or Touch-Tone pad on a telephone. It is *not* the final product. Therefore, speech technologies must be *hidden* from the user. That is, the burden of using the technology must be on the technology itself. (*iii*) *Quick prototyping and development of new products and services.* The technology must support the creation of new products and services based on speech in an efficient and timely fashion. In this paper I present a vision of the voice-processing industry with a focus on the areas with the broadest base of user penetration: speech recognition, text-to-speech synthesis, natural language processing, and speaker recognition technologies. The current and future applications of these technologies in the telecommunications industry will be examined in terms of their strengths, limitations, and the degree to which user needs have been or have yet to be met. Although noteworthy gains have been made in areas with potentially small user bases and in the more mature speech-coding technologies, these subjects are outside the scope of this paper.**

As the telecommunications industry evolves over the next decade to provide the products and services that people will desire, several key technologies will become commonplace. Two of these, automatic speech recognition (ASR) and text-to-speech synthesis (TTS), will provide users with more freedom regarding when, where, and how they can access information. Although these technologies are currently in their infancy, their capabilities are increasing rapidly and their use in today's telephone network is expanding.

Beginning with advances in speech coding, which now allows for high-speed transmission of audio signals, speech-processing technologies and telecommunications are the perfect marriage of a technology and an industry. Currently, the voice-processing market is projected to be over $1.5 billion by 1994 and is growing at about 30% a year (1–3). The two driving forces behind this growth are (*i*) the increased demand for interactive voice services such as voice response and voice messaging and (*ii*) the rapid improvement in speech recognition and synthesis technologies.

Current applications using speech recognition and text-to-speech synthesis technologies center around two areas: those that provide cost reduction [e.g., AT&T and Bell Northern Research's (BNR) automation of some operator functions and NYNEX and BNR's attempt to automate portions of directory assistance] and those that provide for new revenue opportunities [e.g., AT&T's Who's Calling service, NYNEX's directory assistance call completion service, BNR's stock quotation service, and Nippon Telegraph & Telephone's (NTT) banking by phone service].

Yet in the face of this potentially large market, a quarter century ago the influential John Pierce wrote an article questioning the prospects of one technology, speech recognition, and criticizing the "mad inventors and unreliable engineers" working in the field. In his article entitled "Whither speech recognition," Pierce argued that speech recognition was futile because the task of speech understanding is too difficult for any machine (4). Such a speech-understanding system would require tremendous advances in linguistics, natural language, and knowledge of everyday human experiences. In this prediction he was completely correct: there is still no speech recognizer that can transcribe natural speech as well as a trained stenographer, because no machine has the required knowledge and experience of human language. Furthermore, this ultimate goal is still not within sight today. Pierce went on to describe the motivation for speech recognition research: "The attraction [of speech recognition] is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon." His influential article was successful in curtailing, but not stopping, speech recognition research.

What Pierce's article failed to foretell was that even limited success in speech recognition—simple, small-vocabulary speech recognizers—would have interesting and important applications, especially within the telecommunications industry. In 1980 George Doddington, in another "Whither speech recognition?" article, pointed this out (5). He emphasized that it was unnecessary to build the ultimate speech-understanding system with full human capabilities to get simple information over the telephone or to give commands to personal computers. In the decade since Doddington's article, tens of thousands of these "limited" speech recognition systems have been put into use, and we now see the beginnings of a telecommunications-based speech recognition industry (2, 6–10). The economic impact of just one application, the automation of operator services, is well over $100 million a

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

year. However, there are still many technical challenges that must be surmounted before universal use of ASR and TTS technologies can be achieved in the telephone network. These challenges include:

- *High level of accuracy*. The technology must be perceived by the user as highly accurate, robust, and reliable. Voice-processing systems must operate under various conditions—from quiet living rooms to noisy airport terminals—while maintaining high performance levels for all conditions. Over time, as a worldwide wireless telecommunications network becomes a reality, speech technology *must* grow to provide the desired interface between the network and the communications services that users will demand.
- *Easy to use*. Speech is only one of several possible input/output modalities for conveying information between a human and a machine, much like a computer terminal or Touch-Tone pad on a telephone. It is *not* the final product. Therefore, speech technologies must be *hidden* from the user. That is, the burden of using the technology must be on the technology itself. For example, TTS systems must be natural and pleasant sounding, and ASR systems must be able to recognize predefined vocabulary words even when nonvocabulary words are also uttered.
- *Quick prototyping and development of new products and services*. The technology must support the creation of new product and service ideas based on speech in an efficient and timely fashion. Users should be not required to wait weeks or months for new products or services.

In this paper, I present a vision of the voice-processing industry, with a focus on the areas with the broadest base of user penetration: speech recognition, text-to-speech synthesis, natural language processing, and speaker recognition technologies. Current and future applications of these technologies in the telecommunications industry will be examined in terms of their strengths, limitations, and the degree to which user needs have been or have yet to be met. Basic research is discussed elsewhere in this volume. In this paper, I discuss only the basic technologies as they relate to telecommunications-based applications needs. Although noteworthy gains have been made in areas with potentially small user bases and in the more mature speech-coding technologies, these subjects are outside the scope of this paper.

THE VISION

At AT&T we have developed a vision for voice processing in the telecommunications industry that will carry us into the next century:

To have natural, continuous, two-way communication between humans and machines in any language, so that people have easier access to one another, to information, and to services—anytime and anywhere.

This is a very ambitious vision and one that will take decades to achieve. *Natural, continuous, two-way communication*—speech recognition technology can currently support only small vocabularies spoken in a rather stylized fashion (6, 8, 11, 12), and while a text-to-speech system can produce intelligible speech from practically any text it is presented with, it is anything but natural sounding. *Two-way communication* implies being able to translate speech from one language to another so that people can communicate across language barriers—a tall order for current state-of-the-art techniques (13–17). *So that people have easier access to one another, to information, and to services* implies that we must be able to extract from a speech signal relevant information that can provide a computer with the data it needs to obtain, infer, create, or compute the information desired by the user. We are just beginning to understand how to incorporate natural language processing into the speech recognition world so that the meaning of a user's speech can be extracted. This research is in its infancy and may require more than a decade of work before viable solutions can be found, developed, and deployed (18–20). We are far from having such technology ready for deployment within the telecommunications industry. *Anytime and anywhere*—this, too, is a tall order. Technology must be robust enough to work equally well from the quietest ones (e.g., an office) to the noisiest ones (e.g., an airport or moving car). Users cannot be bothered with having to *think* about whether the technology will work. It either does and will become ubiquitous in society or it does not and will be relegated to niche applications.

Visions like this are what drives the speech community. Someday it will become a reality. It is important to understand that speech technology is *not* the final product. It is only another modality of input and output (much like keyboards and Touch-Tone pads), which will provide humans with an easier, friendlier interface to the services desired. While we wait for our vision to be realized, there are many so-called "low-hanging-fruit" telecommunications-based applications that current speech technologies can support that do not need the full capabilities just described. Many of these are discussed in the sections that follow.

THE ART OF SPEECH RECOGNITION AND SYNTHESIS

Current speech recognition and text-to-speech synthesis practices encompass engineering art as well as scientific knowledge. Fundamental knowledge of speech and basic principles of pattern matching have been essential to the success of speech recognition over the past 25 years. Knowledge of basic linguistics and signal-processing techniques has done the same for synthesis. That said, the *art* of successful engineering is critically important for applications using these technologies. There is an important element of craftsmanship in designing a successful speech recognition or text-to-speech-based application. Knowledge of the task also helps ASR- and TTS-based applications be tuned to the user's requirements. Often, this engineering art is developed through trial and error. It should be emphasized that improving the engineering art is a proper and necessary topic for applied research.

The art of speech recognition and synthesis technology has improved significantly in the past few years, further opening up the range of possible applications (21). For speech recognition some of the advances are:

- *Wordspotting*. We are a long way from understanding fluently spoken spontaneous speech. However, some very simple elements of language understanding have been successfully developed and deployed. The ability to spot key sounds in a phrase is the first step toward understanding the essence of a sentence even if some words are not or cannot be recognized. For example, in the sentence *I'd like to make a collect call please*, the only word that must be recognized in an operator services environment is the key word *collect*. Given that hundreds of millions of potential users will be able to pick up their telephones and use a speech recognizer to perform some task, to assume that the users will strictly adhere to the protocol of speaking only words that the recognizer understands is grossly naive. Thus, wordspotting, or the ability to recognize key words from sentences that contain both key words and nonkey words, is essential for any telecommunications-based application (12, 22–24).
- *"Barge in."* When talking with a person, it is desirable to be able to interrupt the conversation. In most current telephone-based voice response systems, it is possible to interrupt a prompt using Touch-Tones. This capability has been extended to allow users the option to speak during a prompt and have the system recognize them. "Barge in" provides a necessary, easy-to-use capability for customers and, as with word-

spotting, is an essential technology for successful mass deployment of ASR into the telephone network (25).

- *Rejection*. An ability that we take for granted in conversation is the ability to detect when we do not understand what someone is saying. Unfortunately, this is a very difficult task for current speech recognition systems. While it is possible to determine when there are two (or more) possible words or sentences, it has been very difficult for systems to determine when people are saying something on a completely different subject. Given the diversity of customers in the telephone network that would use speech recognition capabilities, accurately rejecting irrelevant input is mandatory. Further research effort is needed in detecting this type of "none of the above" response (12, 22–24).
- *Subword units*. It is now possible to build a speaker-independent dictionary of models comprised of constituent phonetic (or phoneme-like) statistical models. Initially, this work focused on supporting robust speech recognition for small, easily distinguishable vocabularies. More recently the effort has focused on supporting larger-vocabulary applications (9). These subword pieces are then concatenated to build representative models for arbitrary words or phrases. Therefore, the effort and expense of gathering speech from many speakers for each new vocabulary word are eliminated, making the development and deployment of new and improved applications simple, quick, and efficient. Much of this work has relied on the knowledge gained from work in TTS. For example, the rules for describing new words in terms of subword units can be derived from the rules used by TTS systems to allow for proper pronunciation of words or phrases (15, 26).
- *Adaptation*. People can adapt quickly to dialects and accents in speech. It is rather naive to think that we can develop a set of models for a speech recognition system that can recognize all variations in speaking and channel conditions. Machines now have the beginnings of the capability to respond more accurately as they learn an individual voice, dialect, accent, or channel environment (27–29).
- *Noise immunity and channel equalization*. Better speech enhancement algorithms and channel modeling have made speech recognizers more accurate in noisy or changing environments, such as airports or automobiles (30–33).

For text-to-speech synthesis, some advances in the engineering art include:

- *Proper name pronunciation*. In general, proper names do not follow the same prescribed rules for pronunciation as do other words. However, one of the major applications for TTS technology is to say people's names (e.g., directory assistance applications). Most current TTS systems have implemented techniques to determine the etymology of a name first and then pronounce the name given a set of rules based specifically on its origin. Therefore, *Weissman* would be pronounced with a long *i* (as is common in Germany) as opposed to a long *e* as would be common in English (e.g., as in *receive*) (34).
- *Address, date, and number processing*. Addresses, dates, and numbers have different meanings and pronunciations depending on how they are used in an address or sentence. For example, does the abbreviation *St.* stand for *Street* or *Saint*? Is *Dr.* for *Drive* or *Doctor*? And what happens if no punctuation is provided with the text, in which case *oh* could mean *Ohio*. In the past decade, much work has gone into making TTS systems much more robust to these types of requirements. For specific applications, most current systems have no problems with this type of input. There is ongoing research in text analysis to improve the performance of TTS in the most general cases (35).
- *Prosody*. While a natural-sounding voice is an obvious goal of TTS research, current technology still produces "machine"-sounding voices. However, in the past few years the incorporation of better prosodic modeling, such as pitch, duration, and rhythm, has greatly increased the melodic flow or intonation of the TTS voice (36, 37).

The design of an easy-to-use dialogue with a computer system is a significant challenge. We know from experience that it is possible to design good human interfaces for computer dialogue systems. Unfortunately, it has also been verified that it is possible to design systems that aggravate people. At this time there are some general guidelines for good human interface designs, but there is no "cookbook" recipe that guarantees a pleasant and easy-to-use system (38). Thus, the art of speech recognition and TTS technologies need to be advanced while waiting for major research breakthroughs to occur.

APPLICATIONS OF SPEECH RECOGNITION AND SYNTHESIS

It is important to bear in mind that the speech technologies described above, notwithstanding advances in reliability, remain error-prone. For this reason the first successful products and services will be those that have the following characteristics:

- *Simplicity*. Successful services will be natural to use. For instance, they may use speech recognition to provide menu capabilities using only small vocabularies (less than 10 words), rather than large vocabularies (more than 1000 words).
- *Evolutionary growth*. The first applications will be extensions of existing systems—for example, speech recognition as a Touch-Tone replacement for voice response systems or TTS for reading out information stored in a machine, such as for remote electronic mail access.
- *Tolerance of errors*. Given that any speech recognizer and synthesizer will make occasional errors, inconvenience to the user should be minimized. This means that careful design of human factors will be essential in providing suitable systems.

That said, there are some general questions that must be asked when considering an application using current speech technologies. The answers will help determine whether it is advisable or possible to design a quality application using speech technology. Some of these questions include:

- Are the potential users friendly and motivated? If so, they might accept a higher error rate in order to carry out the function they desire.
- What environment will the recognizer be expected to work in (e.g., a noisy airport or quiet home)?
- How robust is the algorithm performance in the face of adverse conditions?
- Has the speech technology been prototyped or is it still in the laboratory?
- Can the technology be "broken" by malicious users or hackers?
- How well thought out is the user interface to the technology?
- What accuracy will the user of this service expect?
- What is the maximum tolerable error rate?
- Are the ASR and TTS algorithms accurate enough to meet user expectations?
- Is natural-sounding speech required for the application?
- Does the benefit of using speech technology in this application outweigh its cost compared to alternative technologies?

SPEECH TECHNOLOGY TELECOMMUNICATIONS MARKET

How do the speech technologies described above expand to telecommunications-based products and services? Fig. 1 graphically shows the main application areas for speech recognition, speaker recognition, natural language processing, and text-to-speech synthesis currently considered industry-

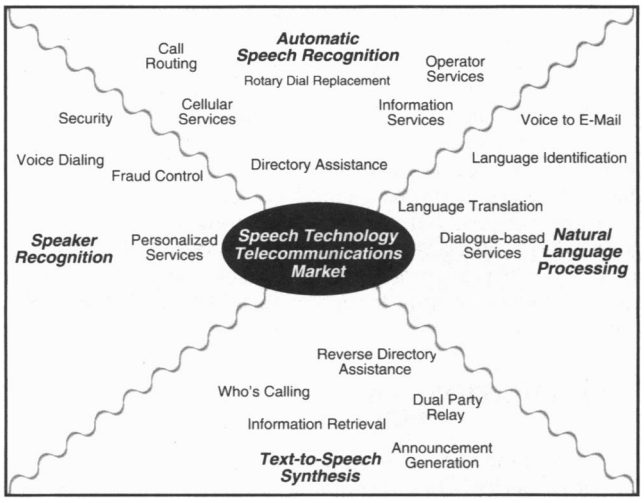

FIG. 1. Figure showing the major application groups for speech-processing technologies in the telecommunications market.

wide. The figure shows that most of the broad application areas center around speech recognition, such as for menu-based transactions or for information access. The fuzzy lines indicate where overlapping technologies are needed. Applications in this area include the whole field of language translation and identification, where the interaction between natural language processing and speech recognition is essential. In the following sections, I will discuss many of the applications currently being deployed, trialed, or planned that fall into these different technology areas. Table 1 gives a summary of all the applications discussed below.

Cost Reduction vs. New Revenue Opportunities

There are two classes of applications that are beginning to appear. The first, *cost reduction applications*, are those for which a person is currently trying to accomplish a task by talking with a human attendant. In such applications the accuracy and efficiency of the computer system that replaces the attendant are of paramount concern. This is because the benefits of ASR technology generally reside with the corporation that is reducing its costs and not necessarily with the end users. Hence, users may not be sympathetic to technology failures. Examples of such applications include (*i*) automation of operator services, currently being deployed by many telephone companies, including AT&T, Northern Telecom, Ameritech, and Bell Atlantic; (*ii*) automation of directory assistance, currently being trialed by NYNEX and Northern Telecom; and (*iii*) control of network fraud currently being developed by Texas Instruments (TI), Sprint, and AT&T.

The second class of applications are services that generate new revenues. For these applications the benefit of speech recognition technology generally resides with the end user. Hence, users may be more tolerant of the technology's limitations. This results in a win-win situation. The users are happy, and the service providers are happy. Examples include (*i*) automation of banking services using ASR offered since 1981 by NTT in Japan; (*ii*) Touch-Tone and rotary phone replacement with ASR introduced by AT&T in 1992; (*iii*) reverse directory assistance, in which a user enters a telephone number to retrieve a name and address provided by Ameritech and Bellcore beginning in 1992; and (*iv*) information access services, such as a stock quotation service currently being trialed by BNR.

In general, the most desirable applications are those that are not gimmicks but provide real usefulness to customers. Since these technologies are in their infancy and have obvious limitations, careful planning and deployment of services must be achieved if mass deployment of the technologies is to occur.

Automation of Operator Services

In 1985 AT&T began investigating the possibility of using limited-vocabulary, speaker-independent, speech recognition capabilities to automate a portion of calls currently handled by operators. The introduction of such a service would reduce operator workload while greatly increasing the overall efficiency of operator-handled calls. The exact task studied was automation of billing functions: *collect*, *calling card*, *person-to-person*, and *bill-to-third-number*. Customers would be asked to identify verbally the type of call they wished to make without speaking directly to a human operator. Could a simple five-word vocabulary (the function names and the word *operator* for human assistance) be designed, built, and deployed with such a degree of accuracy that customers would be willing to use the technology? Early trials in 1986 and 1987 seemed to indicate

Table 1. Network-based, voice-processing-based services

Name	Company	Date	Technology	Task
ANSER	NTT	1981	Small-vocabulary, isolated-word ASR	Banking services
Automated Alternative Billing Services	BNR/Ameritech	1989	Small-vocabulary, isolated-word ASR	Automation of operator services
Intelligent Network	AT&T	1991	Small-vocabulary, wordspotting, barge-in, Spanish	Rotary telephone replacement in Spain
Voice Recognition Call Processing (VRCP)	AT&T	1991	Small-vocabulary, wordspotting, barge-in	Automation of operator services
Telephone Relay Services (TRS)	AT&T	1992	TTS	Enhancement of services to the hearing impaired
Voice Internative Phone (VIP)	AT&T	1992	Small-vocabulary, wordspotting, barge-in	Automated access to enhancement telephone features
Directory Assistance Call Completion (DACC)	NYNEX, AT&T	1992–1993	Small-vocabulary ASR	Automation of directory services
Flex-Word	BNR	1992	Large-vocabulary isolated-word ASR	Stock quotations and automation of directory assistance
Reverse Directory Assistance (RDA)	Bellcore/Ameritech	1993	TTS	Name and address retrieval
Voice Dialing	NYNEX	1993	Small-vocabulary, speaker-dependent, isolated-word, ASR	Automatic name dialing
Voice Prompter	AT&T	1993	Small-vocabulary, wordspotting, barge-in	Rotary telephone replacement

that the technology was indeed providing such performance levels (39).

In 1989 BNR began deploying AABS (Automated Alternate Billing Services) (11) through local telephone companies in the United States, with Ameritech being the first. For this service, ASR and Touch-Tone technologies are used to automate only the back end of collect and bill-to-third-number calls. That is, after the customer places a call, a speech recognition device is used to recognize the called party's response to the question: *You have a collect call. Please say yes to accept the charges or no to refuse the charges.* Using the two-word recognition system (with several yes/no synonyms), the system successfully automated about 95% of the calls that were candidates for automation by speech recognition (6).

After extensive field trials in Dallas, Seattle, and Jacksonville during 1991 and 1992, AT&T announced that it would begin deploying VRCP (Voice Recognition Call Processing). This service would automate the front end as well as the back end of collect, calling card, person-to-person and bill-to-third-number calls. These trials were considered successful not just from a technological point of view but also because customers were willing to use the service (7). By the end of 1993, it is estimated that over 1 billion telephone calls each year will be automated by the VRCP service.

What differentiates the earlier BNR system from the AT&T system is the speech recognition technology. Analysis of the 1985 AT&T trials indicated that about 20% of user utterances contained not only the required command word but also extraneous sounds that ranged from background noise to groups of nonvocabulary words (e.g., "I want to make a *collect* call please"). These extraneous sounds violated a basic assumption for many speech recognition systems of that time: that the speech to be recognized consist solely of words from a predefined vocabulary. With these constraints, the burden of speaking correctly fell on users. In 1990 AT&T developed its wordspotting technology, which began to shift the burden from the users to the speech recognition algorithms themselves. This technology is capable of recognizing key words from a vocabulary list spoken in an unconstrained fashion (12). Results from field trials showed that about 95% of the calls that were candidates for automation with speech recognition were successfully automated when wordspotting was used to accommodate all types of user responses.

We expect that the capability to spot key words in speech will be a prerequisite for most telephone network applications. Also, the ability to recognize speech spoken over voice prompts, called *barge in*, is essential for mass deployment of ASR technology in the network (25). Both of these techniques have been deployed in VRCP.

In 1992, Northern Telecom announced (9) a trial for the automation of a second operator function, directory assistance. This service would rely on technology that the company calls *Flexible Vocabulary Recognition*. By entering the pronunciation of words in a more basic phonetic-like form, pattern-matching methods can be used to find sequences of subword units that match sequences in the pronunciation "dictionary." Thus, vocabularies of hundreds or thousands of words can, in principle, be recognized without having to record each word. This is especially convenient for vocabularies for which new words need to be added when the service is already in use, for instance, names in a telephone directory.

The directory assistance service would allow telephone customers to obtain telephone numbers via speech recognition, and only in difficult cases would a human operator be necessary. As currently envisioned, the customer would place a call to directory assistance, and hear a digitized voice asking whether the caller preferred French or English (the initial service is planned for Canada). After the caller says "English" or "Francais," subsequent speech prompts would be given in that language. Next the caller would be asked to say the name of the city. The customer's response, one of hundreds of city names in Canada, would be recognized using speaker-independent word recognition based on subword units. The caller would then be transferred to an operator, who would have the information spoken thus far displayed on a screen. Subsequent stages of the call—for instance, recognition of the names of major businesses—would be automated in later phases of the trial deployment.

Voice Access to Information over the Telephone Network

It has been over a decade since the first widespread use of automatic speech recognition in the telephone network was deployed. In 1981 NTT combined speech recognition and synthesis technologies in a telephone information system called *Anser*—Automatic Answer Network System for Electrical Requests (40). This system provides telephone-based information services for the Japanese banking industry. Anser is deployed in more than 70 cities across Japan serving over 600 banks. Currently, over 360 million calls a year are automatically processed through Anser, bringing in about $30 million in revenue to NTT annually.

Using a 16-word lexicon consisting of the 10 Japanese digits and six control words, a speaker-independent, isolated-word, speech recognition system allows customers to make inquires and obtain information through a well-structured dialogue with a computer over a standard telephone. Currently, about 25% of customers choose to use the ASR capabilities, with a reported recognition accuracy of 96.5% (41).

Anser provides a win-win scenario for both the service provider and the customer. From a customer's standpoint, the cost of obtaining information about bank accounts is low (about the cost of a local telephone call). Also, because most banks are on the Anser network, there is a uniformity across the banking industry. Therefore, customers can access any bank computer using the same procedures anytime of the day. For the banking industry, Anser allows the banks to provide a much-needed service to its customers without having to hire large numbers of people or invest heavily in equipment.

Anser also became a successful service because the banks demanded that it be accessible from any ordinary telephone. In 1981 more than 80% of the telephones in Japan were rotary dial. Even as late as 1990 more than 70% were still rotary. Therefore, speech recognition was an essential technology if Anser was to be successful in the marketplace.

An emerging area of telephone-based speech-processing applications is that of Intelligent Networks services. AT&T currently offers network-based services via a combination of distributed intelligence and out-of-band common channel signaling. For example, a current 800 service might begin with the prompt *Dial 1 for sales information or 2 for customer service.* Until now, caller input to Intelligent Network services required the use of Touch-Tone phones. Automatic speech recognition technology has been introduced to modernize the user interface, especially when rotary phones are used. Using the AT&T 800 Speech Recognition service, announced at the beginning of 1993, menu-driven 800 services can now respond to spoken digits in place of Touch-Tones and will provide automatic call routing.

Intelligent Network services have attracted the interest of international customers. The first AT&T deployment with speech recognition was in Spain in 1991 (8). The low Touch-Tone penetration rate in Spain (less than 5%) and the high-tech profile of Telefónica, the Spanish telephone company, were the primary motivating factors. In a sense, Telefónica is trying to leap-frog past Touch-Tone technology with more advanced technologies. The targets of this application were conservatively set to accommodate the Spanish telephone network and its unfamiliar users. The speech recognizer deployed supports speaker-independent isolated word recog-

nition with wordspotting and barge in of the Spanish key words *uno*, *dos*, and *tres*. Fig. 2 illustrates an information service based on this platform. With this system a user can obtain information on any of nine topics with just two voice commands.

Particular attention was paid to recognition of key words embedded in unconstrained speech. For isolated words the recognizer correctly identifies the word 98% of the time. Considering all customer responses, including words embedded in extraneous speech, an overall recognition accuracy of 95% or better is expected. Similar systems were also tested in England, Scotland, Wales, and Ireland in 1991 (8).

Subword-based speech recognition has recently been applied to an information access system by BNR working with Northern Telecom (9). Using the 2000 company names on the New York Stock Exchange, callers to this system can obtain the current price of a stock simply by speaking the name of the stock. Though the accuracy is not perfect, it is adequate considering the possibilities of confusion on such a large vocabulary. The experimental system is freely accessible on an 800 number, and it has been heavily used since its inception in mid-1992. There are thousands of calls to the system each day, and evidence suggests that almost all callers are able to obtain the stock information they desire.

The general concept is voice access to information over the telephone. People obtain information from computer databases by asking for what they want using their telephone, not by talking with another person or typing commands at a computer keyboard. As this technology develops, the voice response industry will expand to include voice access to information services such as weather, traffic reports, news, and sports scores. The voice response services will begin to compete with "news radio" and computer information networks. Soon people who get information from computer databases by telephone will not think of themselves as sophisticated computer users. Ease of use is the key; people will think of themselves as having a brief conversation with a machine to get the information they need.

Voice Dialing

One of the biggest new revenue-generating applications of speech recognition technology in telecommunications is voice dialing, or a so-called *Voice Roledex*. Currently, we must all remember hundreds of phone numbers of people and businesses that we need to call. If we want to call someone new, we either look the number up in a phone book or call directory assistance to get the number. But the phone number is only a means to an end—the end being that we want to place a phone call. The fact of the matter is that people really do not want to keep track of phone numbers at all. We should be able to just speak the name of the party that we want to reach and have the phone call be placed totally automatically. This is one example of the *people have easier access to one another* part of our vision.

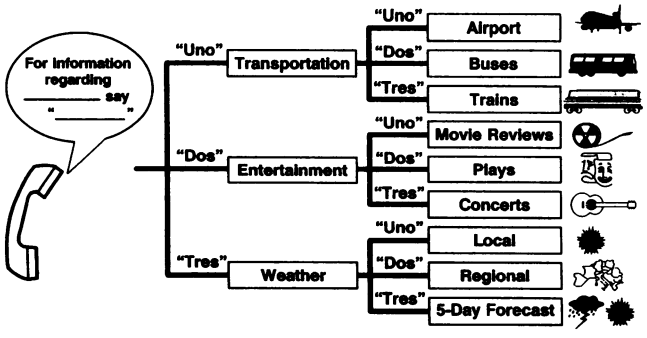

FIG. 2. Example of an ASR-based service using AT&T's Intelligent Network.

Obviously, current ASR technology is not advanced enough to handle such requests as, *Please get me the pizza place on the corner of 1st and Main, I think it's called Mom's or Tom's.* However, most requests to place telephone calls are much simpler than that—for example, *Call Diane Smith*, *Call home*, or *I'd like to call John Doe.* ASR technology, enhanced with wordspotting, can easily provide the necessary capabilities to automate much of the current dialing that occurs today.

Voice-controlled repertory dialing telephones have been on the market for several years and have achieved some level of market success. Recently, NYNEX announced the first network-based voice dialing service, whereby the user picks up his or her home phone and says the name he or she would like to call. It is a speaker-trained system supporting about 50 different names and does not have wordspotting capabilities. NYNEX believes this service will be a market winner, expecting over 10% of its customers to enroll.

One of the main drawbacks of the NYNEX system is that it is tied to a person's home or office telephone. Customers cannot use the service while they are away from their base phone. Currently, AT&T, Sprint, MCI, and TI are trialing expanded versions of the voice-dialing application that handles the large *away from home and office* market. These systems allow users to place phone calls using speech recognition from any phone, anywhere. In addition, these systems also plan to use speaker verification technology to provide a level of network security for the user. Users would have to enter a voice password or account number that would be used to verify their identity before allowing them to place a call. The AT&T system will also use speaker-independent, subword-based ASR instead of speaker-trained ASR for the name-dialing function of the service. This will provide an enormous reduction in data storage and will allow TTS technology to be used to read back the name that is being called. These systems will be available during 1994.

Telephone Relay Service

For AT&T's Telephone Relay Service (TRS), text-to-speech synthesis is used to help hearing-impaired individuals carry on conversations with normal-hearing individuals over telephone lines by minimizing the need for third-party intervention or eliminating the need for both parties to have TDD (Terminal Device for the Deaf) terminals. It is assumed that one party is hearing impaired and has a TDD terminal and the other party has no hearing impediment and no special terminal device.

After dialing into the TRS service, an operator is assigned to the call. As the hearing party speaks, the operator transcribes the speech on a terminal. (Obviously, one would like to have a speech recognizer listening to the incoming speech. However, as stated earlier, ASR technology currently cannot support recognition of fluent spontaneous speech spoken by anyone on any topic.) The text is then transmitted to the hearing-impaired party's TDD unit. When the hearing-impaired party enters a response on his or her TDD, that text is transferred to a TTS system, which then plays out his or her response to the hearing party. This would allow anyone to communicate with a hearing-impaired person without a TDD device. It should be noted that this service has existed without TTS for several years.

The TRS service with TTS technology was trialed by AT&T in California in 1990. Fifteen operator positions were equipped with TTS equipment, and 15 positions were staffed by live operators (as a control) who would read the text to the hearing party. Over 35,000 calls were processed by TTS. Greater than 80% of the calls successfully used TTS for the entire duration of the call. Eighty-eight percent of TDD customers and 86% of hearing customers rated the TTS service as good or

excellent. The worst problem was incorrect handling of spelling errors. Currently, this service is being deployed throughout Washington state.

There were many technical challenges that stood in the way of automating this service using TTS technology (ref. 42; J. Tschirgi, personal communication, 1993). For example, people using TDDs:

- Use only single-case type, usually uppercase. Since TTS systems generally make use of upper- and lowercase information in determining pronunciation rules, modifications had to be made in order to handle this type of text input.
- Do not use punctuation. Therefore, there are no sentence boundaries. Additionally, there is no way to disambiguate whether a sequence of letters is an abbreviation or an actual word.
- Use special abbreviations and contractions, for example *XOXOXO* for *love and kisses*; *OIC* for *Oh, I see*; *PLS* for *please*; and *Q* to indicate a question.
- Use regional abbreviations depending on where they live, for example, LAX for Los Angeles Airport.
- Misspell about 5% of words they type. Obviously, this will cause problems for any TTS system.

All of these issues required extensive research and development before a successful service was deployed.

FUTURE POSSIBILITIES

It has been observed that predictions of future technologies tend to be overly optimistic for the short term and overly pessimistic for the long haul. Such forecasts can have the unfortunate effect of creating unrealistic expectations leading to useless products, followed by premature abandonment of the effort. I have tried to counteract this tendency by carefully pointing out the limitations of current speech recognition and text-to-speech technologies while focusing on the types of applications that can be successfully deployed for mass user consumption given today's state of the art.

Near-Term Technical Challenges

While the prospect of having a machine that humans can converse with as fluently as they do with other humans remains the Holy Grail of speech technologists and one that we may not see realized for another generation or two, there are many critical technical problems that I believe we will see overcome in the next 2 to 5 years. Solving these challenges will lead to the ubiquitous use of speech recognition and synthesis technologies within the telecommunications industry. The only question is how these advances will be achieved. For speech recognition, these research challenges include:

- *Better handling of the varied channel and microphone conditions*. The telephone network is constantly changing, most recently moving from analog to digital circuitry and from the old-style nonlinear carbon button-type transducers to the newer linear electret type. Each of these changes affects the spectral characteristics of the speech signal. Current ASR and especially speaker verification algorithms have been shown to be not very robust to such variability. A representation of the speech signal that is invariant to network variations needs to be pursued.
- *Better noise immunity*. While advances have been made over the past few years, we are a long way away from recognition systems that work equally well from a quiet home or office to the noisy environments encountered at an airport or in a moving car.
- *Better decision criteria*. For a long time, researchers have mainly considered the speech recognition task as a two-class problem, either the recognizer is right or it is wrong, when in reality it is a three-class problem. The third possibility is that of a *nondecision*. Over the past few years, researchers have begun to study the fundamental principles that underlie most of today's algorithms with an eye toward developing the necessary metrics that will feed the creation of robust rejection criteria.
- *Better out-of-vocabulary rejection*. While current wordspotting techniques do an excellent job of rejecting much of the out-of- vocabulary signals that are seen in today's applications, they are by no means perfect. Since AT&T announced that its wordspotting technology was available for small-vocabulary applications in its products and services beginning in 1991, many other ASR vendors have realized that the ability to distinguish key word from nonkey word signals is mandatory if mass deployment and acceptance of ASR technology are to occur. Hence, more and more ASR products today are being offered with wordspotting capabilities. Additionally, as our basic research into more advanced, large-vocabulary systems progresses, better out-of-vocabulary rejection will continue to be a focusing point. With all this activity being directed to the issue, I am sure we will see a steady stream of new ideas aimed at solving this problem.
- *Better understanding and incorporation of task syntax and semantics and human interface design into speech recognition systems*. This will be essential if we are to overcome the short-term deficiencies in the basic technologies. As ASR-based applications continue to be deployed, the industry is beginning to understand the power that task-specific constraints have on providing useful technology to its customers.

Challenges for text-to-speech synthesis research include:

- *More human-sounding speech*. While totally natural speech is decades away, improvements in prosody and speech production methodology will continue to improve the quality of the voices we hear today. One point to note: there are only a handful of laboratories currently working on TTS research; therefore, advances in TTS technology will most likely occur at a slower pace than those made in speech recognition.
- *Easy generation of new voices, dialects, and languages*. Currently, it takes many months to create new voices or to modify existing ones. As more personal telecommunications services are offered to customers, the ability to customize voices will become very important. A simple example of this might be the reading of e-mail. If I know that the e-mail was sent by a man or woman (or a child), the synthesizer should be able to read the text accordingly.

Personal Communication Networks and Services

One of the most exciting new uses of speech technologies is in the area of Personal Communication Networks (PCNs) and Personal Communication Services (PCSs). It is quite obvious that as Personal Communication Devices (PCDs) come of age, their very nature will require the use of advanced speech technologies. As these devices become smaller and smaller, there will be no room for conventional Touch-Tone keypads or any other type of mechanical input device. What room will be available will undoubtedly be used for a video display of some kind. Moreover, the display will more than likely be too small for touch screen technologies other than those that use pen-based input. Thus, speech technologies will become necessary if we are to easily communicate with our personal communicators.

Within the next 2 to 3 years I expect to see some rudimentary speech recognition technology incorporated into PCDs. Initially, ASR will provide users with simple menus for maneuvering around the PCD, including the ability to place telephone calls across the wireless network. Voice response technology will also be included to provide audio feedback to users. This will most probably be done by using current voice coding techniques and then migrating to TTS as the technology becomes implementable on single chips and the large memory requirements of current TTS techniques can be reduced.

Predictions

Predicting a generation in the future may be a futile exercise. It is impossible to predict when a technical revolution will occur; few people could have predicted in the 1960s the impact that VLSI would have on our society. There is also the risk of being blinded by the past when looking to the future. All we can say with assurance is that the present course of our technology will take us somewhat further; there are still engineering improvements that can be built on today's science. We can also anticipate, but cannot promise, advances in scientific knowledge that will create the basis upon which a *new* generation of speech recognizers and synthesizers will be designed.

Let me go out on a limb (a bit) and make some specific predictions:

- Algorithms for highly accurate, speaker-independent recognition of large vocabularies will soon become available. Before the year 2000, this technology will be successfully engineered into specific large-scale applications that are highly structured, even if the vocabulary is large.
- Major advances will be made in language modeling for use in conjunction with speech recognition. In contrast to the past two decades, in which advances were made in feature analysis and pattern comparison, the coming decade will be the period in which computational linguistics makes a definitive contribution to "natural" voice interactions. The first manifestations of these better language models will be in *restricted*-domain applications for which specific semantic information is available, for example, an airline reservation task (18–20, 43).
- Despite the advances in language modeling, the speech-understanding capability of computers will remain far short of human capabilities until well into the next century. Applications that depend on language understanding for *unrestricted* vocabularies and tasks will remain a formidable challenge and will not be successfully deployed for mass consumption in a telecommunications environment for several decades.
- Speech recognition over telephone lines will continue to be the most important market segment of the voice-processing industry, both in terms of the number of users of this technology and its economic impact. The ability to get information remotely, either over telephone lines or wireless personal communications systems, will drive many applications and technological advances.
- The introduction of a voice *Intelligent Agent* will occur within the next few years. Users will initially be able to *ask* their agents to perform simple tasks such as calling people, managing messages, and getting information. They will be able to define the personality they want their agent to have—e.g., a choice between a friendly, informal voice or a more formal voice. The availability of these agents will radically change consumer impressions of speech technologies and will lead to a large increase in the number and scope of speech based applications.
- "Simple" applications of speech recognition will become commonplace. By the year 2000, more people will get remote information via voice dialogues than will by typing commands on Touch-Tone keypads to access remote databases. These information access applications will begin as highly structured dialogues and will be specific to narrow domains such as weather information or directory assistance.
- Truly human-quality text-to-speech synthesis technology will not be available for another decade. As is the case for totally unrestricted-vocabulary ASR algorithms, researchers will have to totally rethink the problem in order to achieve our vision.
- Finally, I confidently predict that at least one of the above six predictions will turn out to have been incorrect.

One thing is very clear: sooner than we might expect, applications based on speech recognition and synthesis technologies will touch the lives of every one of us.

1. Meisel, W., ed. (1993) *Speech Recognition Update* (TMA Associates, Encino, CA).
2. Oberteuffer, J., ed. (1993) *ASR News* (Voice Information Associates, Lexington, MA).
3. The Yankee Group (1991) *Voice Processing: The Second Generation of Equipment and Services.*
4. Pierce, J. R. (1969) *J. Acoust. Soc. Am.* **46,** 1029–1051.
5. Doddington, G. R. (1980) in *Trends in Speech Recognition,* ed. Lea, W. (Prentice–Hall, Englewood Cliffs, NJ).
6. Bossemeyer, R. W. & Schwab, E. C. (1991) *Speech Tech. Mag.,* 24–30.
7. Franco, V. (1993) in *Proceedings of the Voice '93 Conference* (San Diego).
8. Jacobs, T. E., Sukkar, R. A. & Burke, E. R. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (Piscataway, NJ).
9. Lennig, M. (1992) in *Proceedings of the First IEEE Workshop on Interactive Voice Technology for Telecommunications Applications* (Piscataway, NJ).
10. Yashchin, D., Basson, S., Kalyanswamy, A. & Silverman, K. (1992) in *Proceedings of the AVIOS.*
11. Lennig, M. (1990) *Computer* **23,** 35–41.
12. Wilpon, J. G., Rabiner, L. R., Lee, C. H. & Goldman, E. R. (1990) *IEEE Trans. Acoust., Speech, Signal Process.* **38,** 1870–1878.
13. Hutchins, W. J. & Somers, H. L. (1992) *An Introduction to Machine Translation* (Academic, New York).
14. Morimoto, T., Iida, H., Kurematsu, A., Shikano, K. & Aizawa, T. (1990) in *Proceedings of the Info JAPAN '90: International Conference of the Information Processing Society of Japan,* pp. 553–559.
15. Rabiner, L. R. & Juang, B. H. (1993) *Fundamentals of Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).
16. Roe, D. B., *et al.* (1992) *Speech Commun.* **11,** 311–319.
17. Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A. & Tebelskis, J. (1991) in *Proc. ICASSP '91,* pp. 793–796.
18. Hirschman, L., *et al.* (1992) in *Proceedings of the DARPA Speech and Natural Language Workshop* (Harriman, NY), pp. 7–14.
19. Marcus, M., ed. (1992) in *Proceedings of the Fifth DARPA Speech and Natural Language Workshop* (Kaufmann, San Mateo, CA).
20. (1993) in *Proceedings of the DARPA Speech and Natural Language Workshop* (Harriman, NY).
21. Roe, D. & Wilpon, J. (1993) *IEEE Trans. Commun.,* 54–62.
22. Rohlicek, J., Russell, W., Roucos, S. & Gish, H. (1989) in *Proc. ICASSP '89,* pp. 627–630.
23. Rose, R. & Hofstetter, E. (1992) *Proc. ICASSP '92.*
24. Sukkar, R. & Wilpon, J. (1993) in *Proc. ICASSP '93* (Minneapolis) **2,** pp. 451–454.
25. AT&T Conversant Systems (1991) CVIS Product Announcement.
26. Lennig, M., Sharp, D., Kenny, P., Gupta, V. & Precoda, K. (1992) in *Proc. ICSLP-92* (Banff, AB, Canada), pp. 93–96.
27. Lee, C., Lin, C.-H. & Juang, B.-H. (1991) *IEEE Trans.* **39,** 806–814.
28. Rosenberg, A. & Soong, F. (1987) *Comput. Speech Lang.* **22,** 143–157.
29. Schwartz, R., Chow, Y. L. & Kubala, F. (1987) in *Proc. ICASSP '87,* pp. 633–636.
30. Acero, A. (1990) Ph.D. thesis (Carnegie–Mellon Univ., Pittsburgh).
31. Hermansky, H., Morgan, N., Buyya, A. & Kohn, P. (1991) in *Proceedings of Eurospeech '91,* pp. 1367–1370.
32. Hirsch, H., Meyer, P. & Ruehl, H. W. (1991) in *Proceedings of Eurospeech '91,* pp. 413–416.
33. Murveit, H., Butzberger, J. & Weintraub, M. (1992) in *Proceedings of the DARPA Speech and Natural Language Workshop* (Harriman, NY), pp. 280–284.
34. Church, K. (1986) *Proc. ICASSP '86* **4,** 2423–2426.
35. Sproat, R., Hirschberg, J. & Yarowsky, D. (1992) in *Proceedings of the International Conference on Spoken Language Processing* (Banff, AB, Canada).
36. Hirschberg, J. (1990) in *ESCA Workshop on Speech Synthesis* (Autrans, France), pp. 181–184.
37. van Santen, J. P. H. (1995) *Comput. Speech Lang.*, in press.
38. Kamm, C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10031–10037.
39. Wilpon, J. G., DeMarco, D. & Mikkilineni, P. R. (1988) in *Proc. IEEE ICASSP '88,* pp. 55–58.
40. Nakatsu, R. (1990) *Computer* **23,** 43–48.
41. Furui, S. (1992) in *Proceedings of the COST-232 Speech Recognition Workshop,* Rome.
42. Bachenko, J., Daugherty, J. & Fitzpatrick, E. (1992) in *Proceedings of the ACL Conference on Applied NL Processing* (Trente, Italy).
43. Pallet, D. (1991) *Speech Nat. Lang.,* 49–134.