reprinted from

COMPUTER

# Putting Speech Recognition to Work in the Telephone Network

Matthew Lennig

Bell-Northern Research and INRS-Télécommunications

The interactive voice technologies include speech recognition, speaker verification, speech encoding and decoding, and speech synthesis. These technologies provide a way for people to interact verbally with computers. Voice interaction is particularly useful over the telephone because it allows people to communicate directly with computers to perform simple tasks without the need for operators.

Of all the interactive voice technologies, perhaps the most challenging is speech recognition, because of the inherent variability in the way we speak. Speaker-independent speech recognition, in which the computer interacts with people who have not previously "trained" it to their speech characteristics, is more difficult than speaker-dependent recognition, in which the system has been trained to a particular user's speech. When the speech to be recognized is transmitted over the telephone network, further variability is imposed by the varying quality of network connections. On top of these difficulties, the telephone network removes potentially useful information for word discrimination by cutting out spectral energy in the speech signal above about 3,300 Hz and below about 300 Hz.

**The early success of an automated call-handling system using interactive voice technologies foreshadows huge savings for telephone companies and a wealth of new services for consumers.**

Speaker-independent speech recognition was the biggest technical hurdle in the development of Northern Telecom's automated alternate billing service (AABS) for collect calls, third-number-billed calls, and calling-card-billed calls. The AABS system automates a collect call by recording the calling party's name, placing a call to the called party, playing back the calling party's name to the called party, informing the called party that he or she has a collect call from that person, and asking, "Will you pay for the call?" The called party responds *yes* or *no* to the speech recognizer, and the call is completed or not, accordingly.

Before embarking on the AABS project, Bell-Northern Research ran a concept trial in the first half of 1988 with Bell Canada's public customers to gain a better understanding of the real-world behavior of speech recognition technology and to find out how acceptable it would be to the public. The application was a directory of "dial-it" services available in the 514 area code (Montreal), which callers accessed by dialing numbers beginning 9-7-6. Offered to the public free of charge, the directory (in French) interacted with the caller, using stored-speech playback and speech recognition with a total vocabulary of 25 words to provide information on the services available, their telephone numbers, and their prices.

During the trial we found that when input words were within the speech recognizer's vocabulary, the rate of substi-

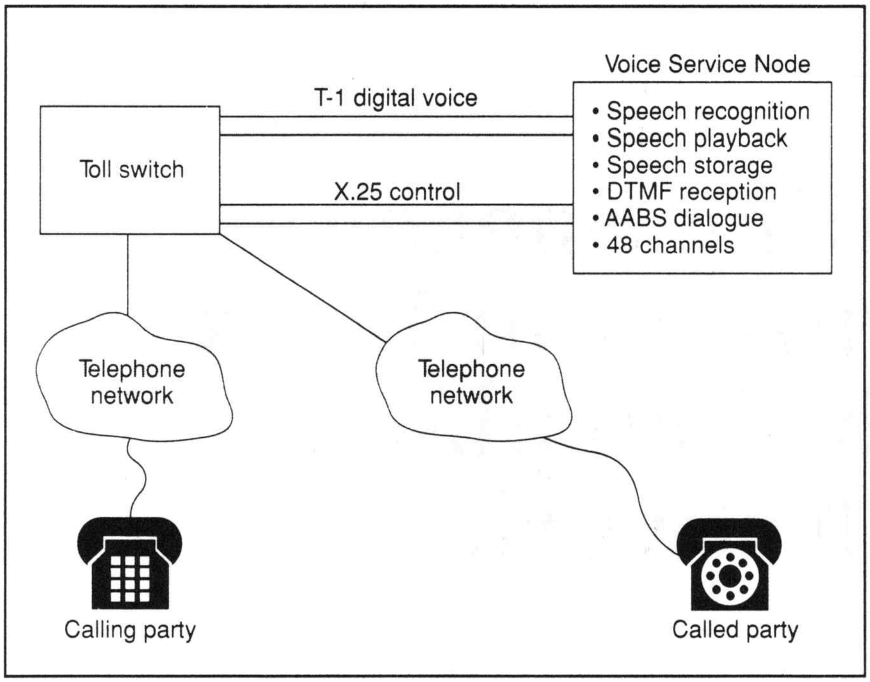

**Figure 1. Network configuration of the Voice Service Node for automation of collect, third-number-billed, and calling card calls.**

tution (recognition of an incorrect word) was 1 percent and the rate of rejection (the input word not accepted as valid) was 3 percent.[1] When input words were from outside the recognizer's vocabulary, a 50-percent false-acceptance rate occurred—that is, the recognizer rejected only half of these invalid words. For the directory application, such a high false-acceptance rate was tolerable, but other applications, such as AABS, cannot function effectively without a much greater discrimination ability.

Both the directory and AABS rely on isolated-word recognition—recognition of a single word or phrase spoken in isolation. In the laboratory, we are currently pursuing more advanced techniques that can recognize continuous speech. For example, we have developed a prototype connected-word recognizer, which is also speaker independent and works over the telephone network.

Certain applications require much larger vocabularies. The University of Quebec's National Scientific Research Institute in Telecommunications (INRS-Télécommunications) has developed a speaker-adaptive isolated-word recognition system capable of recognizing a vocabulary of 86,000 English words with 93-percent accuracy.[2] Because this system is based on phonemes (the smallest distinguishing sound units of speech), a new user need not train it on the entire vocabulary. The system can generalize from a short training script of 100 to 200 sentences (1,000 to 2,000 words) to model the new user's pronunciation of all 86,000 words.

One application of the INRS-Télécommunications recognition system is the Talkwriter, a voice dictation typewriter that enables a user to enter text into a computer by speaking instead of typing. To improve the usability of the Talkwriter, INRS-Télécommunications is enhancing its recognition algorithm to allow continuous-speech input, eliminating the need for quarter-second pauses between dictated words. Earlier work using a 5,000-word vocabulary and continuous speech was done by Bahl, Jelinek, and Mercer.[3] Lee, Hon, and Reddy use context-dependent allophone units to perform speaker-independent, continuous-speech recognition of a 1,000-word vocabulary.[4]

## Automated alternate billing service

AABS is of interest to telephone companies because it offers large potential savings in operator time. In the United States in 1988 there were an estimated 579 million intraLATA* collect calls and 89 million intraLATA third-number-billed calls. An operator's average work time (AWT) for a collect call is 34 seconds. If no verification of billing acceptance is performed, the AWT for a third-number-billed call is 25 seconds; with verification the AWT is 43 seconds. The estimated cost per operator work-second is $0.0103. If we consider only intraLATA collect calls, assuming 85-percent automation and 1988 call volume, the potential annual savings in AWT nationwide is more than $172,000,000. Each of the seven regional holding companies would save about $24,600,000 per year.[5] These figures do not include the additional savings that would be realized from the automation of third-number-billed calls.

The interactive voice technologies used in AABS are implemented on a special-purpose Northern Telecom system called the Voice Service Node (VSN), installed in the central telephone office. This equipment is connected to the toll switch via data links (X.25) and digital voice links (T-1), as illustrated in Figure 1. When a 0+ call (a call dialed with a preceding zero) arrives at the switch, it is sent to the VSN. The VSN plays a "bong" tone followed by spoken instructions (in digitally encoded speech) explaining the billing options. The caller indicates a billing method by using DTMF (dual-tone multifrequency, or touch-tone) signaling, that is, by dialing 1-1 for collect, another telephone number for third-number billing, or a calling card number.

A caller choosing collect or third-number billing is asked to say his or her name, which the VSN captures and stores digitally. The VSN then asks the caller to wait while acceptance of charges is obtained from the billed party. Next, the VSN sends a message to the switch over the X.25 control link, requesting that the switch outpulse the call to the billed party (the same as the called party in the case of collect calls).

For example, suppose Danièle Archer wishes to call Joe's Department Store, collect. She dials zero followed by the telephone number of Joe's Department Store. She hears a bong tone followed by the prompt, "For collect calls, dial 1-1. To

*LATA (local access and transport area): a geographical area roughly the size of an area code calling area over which local telephone operating companies carry transmissions.

charge this call to another number, dial the complete billing number now." (The billing number can be either another phone number or Danièle's calling card number.) She wants to make a collect call, so she dials 1-1. The system says, "Please say your name." She says, "Danièle Archer," and the system records it digitally. The switch then originates a call to Joe's Department Store. When Joe answers, Danièle hears the following dialogue between the VSN and Joe (words in boldface are a digital playback of Danièle's voice):

*Joe:* Hello, Joe's Department Store!

*VSN:* This is Michigan Bell. You have a collect call from **Danièle Archer**. Will you pay for the call?

*Joe:* What did you say?

*VSN:* This is Michigan Bell. You have a collect call from **Danièle Archer**. Please answer the following question *yes* or *no*. Will you pay for the call?

*Joe:* Yeah.

*VSN:* Thank you. Please go ahead.

Note that when Joe responds with a phrase outside the speech recognizer's vocabulary ("What did you say?"), the recognizer correctly rejects it. Dialogue error paths are designed to be as graceful and helpful as possible. Various types of speech recognition errors—rejection, no speech, speech too long, speech too short—generally evoke different dialogue branches designed to guide the user to a successful outcome. We have found that good voice dialogue design is extremely important for the success of an interactive voice application. Dialogue design is an art that benefits from extensive experience combined with a perfectionist mentality.

## Architecture of the Voice Interface

AABS uses the following interactive voice technologies: network-based, speaker-independent speech recognition; real-time digital recording (encoding) of speech; real-time digital playback (decoding) of speech; and detection and reception of DTMF signals. All these functions are performed by Northern Telecom's Voice Interface component of the VSN. Each Voice Interface unit can handle six simultaneous voice channels. Eight such units (with a ninth as a spare) provide the VSN with 48 voice channels. The Voice Interface was first used in the Bell Canada 976 Directory trial.

Each Voice Interface contains four printed-circuit cards: three voice cards and one general-purpose processor card, which acts as an interface to the AABS application software. Each voice card contains an application-specific integrated circuit, used to perform the dynamic time warping required for speech recognition (described in the following section).

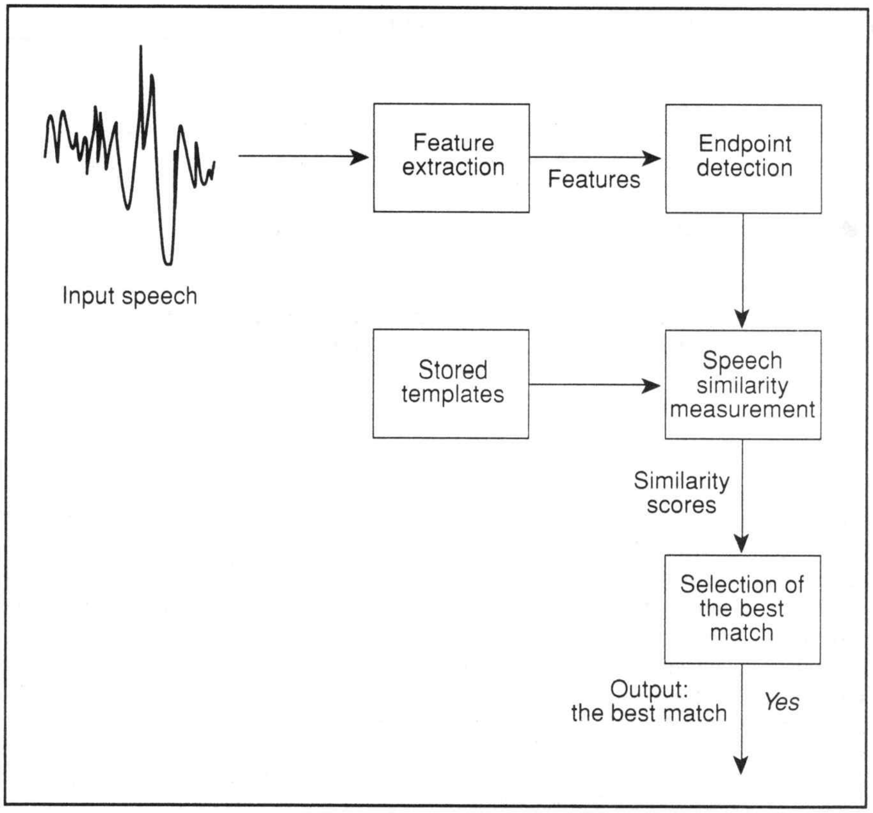

Figure 2. The speech recognizer.

## Speech recognition algorithm

Our speech recognition algorithm involves three main steps: feature extraction, word endpoint detection, and template matching, as illustrated in Figure 2. Feature extraction divides an incoming word into 12.75-millisecond frames and calculates the spectral parameters of each frame. Word endpoint detection locates the beginning and end of the word to be recognized (called the unknown). Pattern matching then compares the spectral properties determined by the feature extraction process to the templates (or stored mathematical models) in the speech recognizer's vocabulary to determine the closest match.

**Feature extraction.** Feature extraction captures the acoustic elements that distinguish one word from another but ignores certain differences in the way the same word is spoken by different callers. The process ignores pitch and absolute loudness, which vary from caller to caller for each word, and concentrates instead on the overall spectral shape and how it changes across the word.

After dividing a word into frames, feature extraction measures the spectral properties of each frame. The sensitivity of the analysis mimics the human ear by exhibiting greater frequency resolution at lower voice frequencies. Telephone transmission limits the range of voice frequencies for analysis. It also requires extra processing to minimize the effect of impairments (for example, additive noise and frequency distortion) introduced by numerous network connections and telephone set types.

The acoustic features used in the current implementation of the speech recognizer

**Table 1. Rules for scoring recognition accuracy.**

| | Recognizer Output | | |
|---|---|---|---|
| Recognizer Input | *Yes* | *No* | Rejection |
| *Yes* | CA | FA | FR |
| *No* | FA | CA | FR |
| Yes-equivalent | CA | FA | CR |
| No-equivalent | FA | CA | CR |
| Imposter | FA | FA | CR |

CA = correct acceptance; FA = false acceptance; FR = false rejection; CR = correct rejection

consist of mel-frequency cepstral coefficients[6] together with their first time differences.[7] The term mel-frequency means that the center frequencies of the channels, instead of being linearly distributed in frequency, are linearly spaced below 1,000 Hz and logarithmically spaced above. This is intended as a rough approximation of the frequency discrimination properties of the ear. The mel-frequency cepstral coefficients are calculated as follows: Every 12.75 ms, a Hamming window of duration 25.6 ms is applied to the input speech signal and a fast Fourier transform is used to compute a power spectrum. Power spectrum points are combined into 20 mel-frequency channels by means of triangular weighting functions. Next, a log transformation is applied to each of the 20 channel energies. Finally, a cosine transform is applied, yielding the mel-frequency cepstral coefficients

$$C_k = \sum_{n=1}^{20} L_n \cos(\pi nk/20 - 1/2)$$

where the $L$'s are the 20 log channel energies, the $C$'s are the cepstral coefficients, and $k$ takes on integer values between 0 and 7.

**Endpoint detection**. To detect the beginning and the end of the unknown word or phrase, endpoint detection distinguishes speech from background noise. It differentiates the two by using a complex set of thresholds and rules to analyze changes in loudness. The thresholds automatically adjust to the diverse signal levels and noise impairments encountered on telephone networks.

Noise poses a significant challenge in endpoint detection because many sounds (such as those produced by the letter *f* and other fricative phonemes) embedded in words resemble telephone noise. Another difficulty is distinguishing the silences contained within words (for example, during the closure of the *p* in *spin*) from the silences that occur at the completion of a word or phrase. The endpoint detector's thresholds and rules not only help differentiate speech from noise, they also distinguish intraword silence from phrase-final silence. The technique used is similar in spirit to that described by Lamel et al.[8]

**Template matching**. The template-matching process is at the heart of the speech recognition technique. It measures the similarity between the unknown and each of the templates in the active vocabulary.

To develop templates for the speech recognizer's vocabulary, we have collected sample words, called tokens, from tens of thousands of English-speaking men and women across the United States and Canada, in a variety of dialects, over a variety of telephone handsets, and over numerous long-distance and local connections. Then we divided the tokens of each word into clusters, each cluster representing similar pronunciations, and one template was generated from each cluster. Together, the templates represent a wide range of pronunciations for a particular word. The speech recognizer compares each unknown against the templates to find the closest match.

During template matching, the template and the unknown must be properly aligned in time before a similarity score can be obtained. The process of time alignment is called dynamic time warping.[9] Dynamic time warping addresses the fact that different speakers not only pronounce a word at different speeds but also elongate different parts of the word. *Dynamic* refers to a technique known as dynamic programming, which determines the optimal piecewise shrinking and stretching of an unknown to match it to a template. *Time warping* is the treatment of the time axes of the unknown and the template as if they were elastic bands—stretching and shrinking different parts of the unknown and the template to maximize the overall similarity score.

Template matching requires extensive computation. Each unknown is compared with the total active vocabulary of the speech recognizer, with from five to 200 templates representing the various pronunciations of each vocabulary word. The top 10 choices go through a second matching process, which uses a different feature extraction strategy—one that is more robust against telephone network impairments but is less precise in discriminating between words. The results of the two strategies are compared, and if they disagree, the result with the larger similarity score ratio is chosen.[10] The similarity score ratio is $s_1/s_2$, where $s_1$ is the similarity score of the top-choice word and $s_2$ is the similarity score of the second highest scoring word.

An application-specific chip, mentioned earlier, cost-effectively and reliably implements the template-matching algorithm. Developed by means of the silicon user design system and CMOS (complementary metal-oxide semiconductor) fabrication technology, the chip performs over 1,000 matches a second.

The outcome of the speech recognition algorithm is an ordered list of matching vocabulary words, with associated similarity scores, for each unknown. The similarity scores indicate the speech recognizer's confidence in a chosen template and help to identify unknowns that are not in its vocabulary.

## Accuracy of the recognizer

Although the recognizer's vocabulary for AABS consists only of the two words *yes* and *no*, under real-world conditions users do not always stay within the constraints of that vocabulary. Potential responses to a yes-or-no question such as *Will you pay for the call?* include *Yes, ma'am*; *Yes, I will*; *Yeah*; *Yup*; *What?*; *Who's this?*; *Hold on a minute*; *No, ma'am*; *Mommy!*; *No way*; *No, thank you*; and many others.

The recognizer can do one of three things with the input utterance: (1) accept

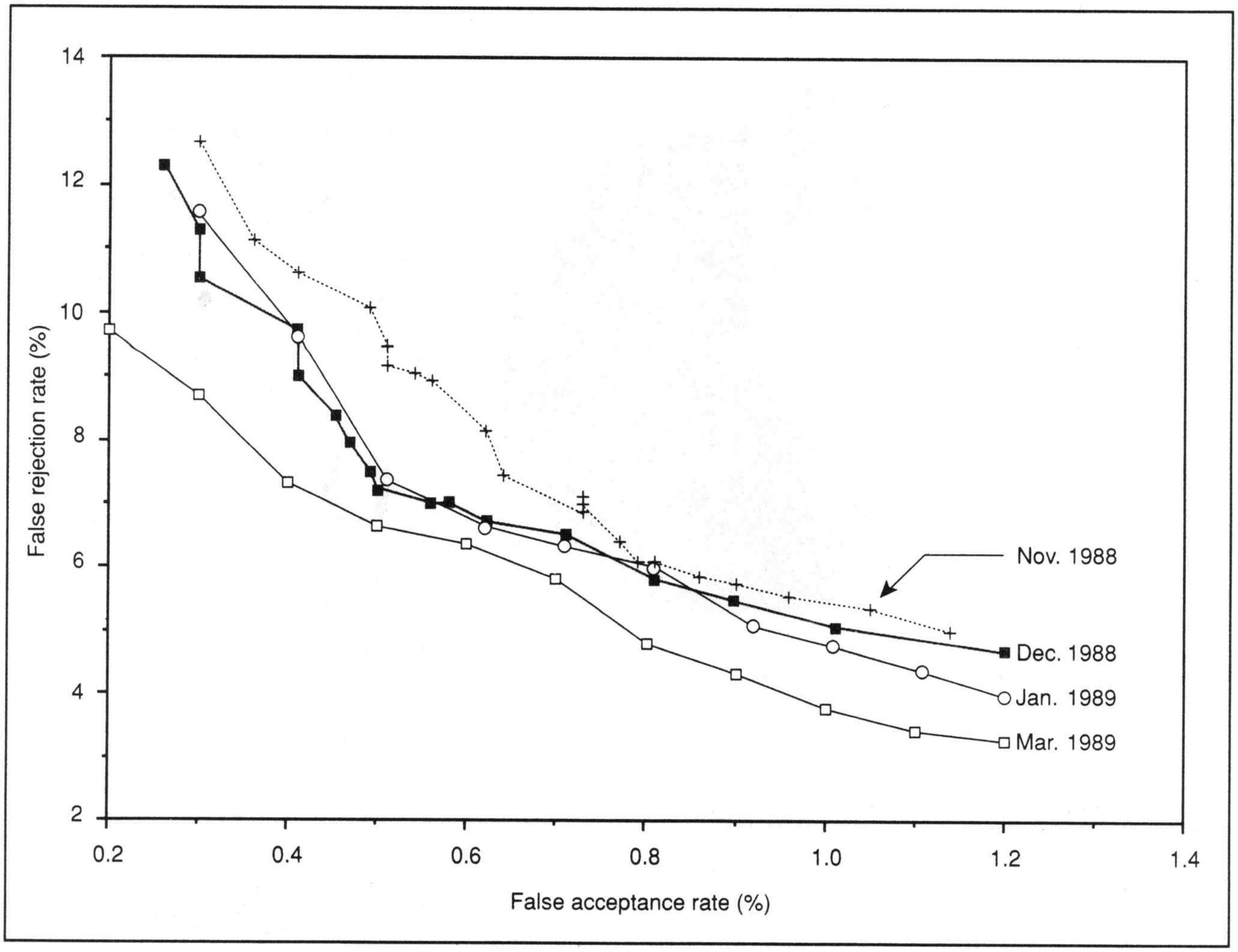

**Figure 3. Trade-off between false acceptance and false rejection for successive refinements of the speech recognizer, based on a 5,021-token test set.**

it as *yes*, (2) accept it as *no*, or (3) reject it as outside the valid vocabulary or as too close to call between the valid choices. When a rejection occurs, the caller is reminded of what the valid input vocabulary is and given a second chance to speak a word from it, as in the dialogue example given earlier.

Four possible outcomes are used to score the recognizer. If the recognizer accepts an input utterance (for example, recognizes it as *no*), the outcome is labeled a correct acceptance (CA) if the classification is correct (that is, the input was *no*) or a false acceptance (FA) if the classification is incorrect (the actual input was *yes* or *who is this?*). If the recognizer rejects the input utterance, the outcome is labeled a correct rejection (CR) if the input was an imposter (a word from outside the vocabulary, such as *what*) or as a false rejection (FR) if the input was a valid *yes* or *no*.

Certain forms of *yes* (such as *yeah, yeh, yuh, yup, ye, yop*) and *no* (*nope, naw, nah, neh*) are considered sufficiently close to the word in question to be mandatory acceptances. Such forms are treated the same as *yes* and *no*. In other words, they are counted as false rejections if they are rejected.

This leaves the problem of what to do with expressions such as *yes, ma'am*, which carry the meaning of *yes* and even have the word embedded in them. We created a special class of such phrases, which we refer to as yes-equivalents. Similarly, no-equivalents are phrases that mean *no* and have a phonetic sequence resembling *no* embedded in them—for example, *no, I will not*.

For scoring purposes, we have chosen to consider yes- and no-equivalents as optionally rejectable. That is, they are counted as correct rejections if rejected and as correct acceptances if correctly accepted (for example, *yes, I will* recognized as *yes*). Table 1 summarizes the scoring rules just described. For example, whenever the recognizer input is *yes* and its output is *no*, the outcome is counted as an FA. Recognition error rates are stated in terms of percent FA and percent FR.

An obvious trade-off exists between FA and FR. The trade-off is controlled through a threshold parameter. Figure 3 shows a family of trade-off curves during the period in which we were fine-tuning the feature extraction, training, and recognition algorithms (November 1988 through March 1989). The test set for each curve consists of 5,021 tokens of *yes*, *no*, and imposters sampled all over the United States and Canada (except Quebec) over

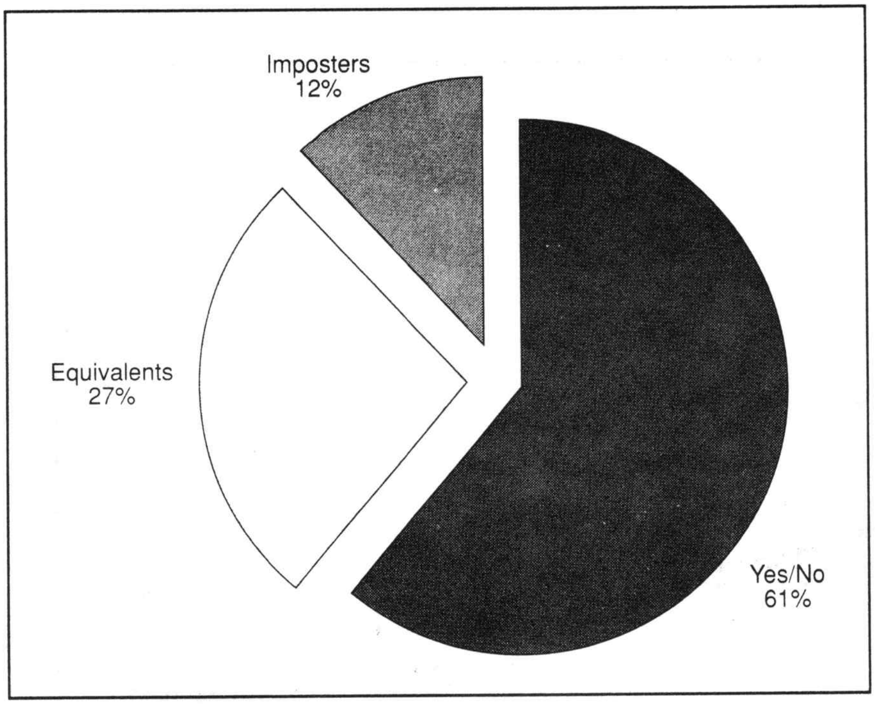

**Figure 4. Composition of the 5,021-token test set.**

**Table 2. Confusion matrix for the March 1989 recognition experiment.**

| | Recognizer Output | | | |
|---|---|---|---|---|
| Recognizer Input | *Yes* | *No* | Rejection | Totals |
| *Yes* | 1,336 | 20 | 83 | 1,439 |
| *No* | 7 | 1,522 | 97 | 1,626 |
| Yes-equivalent | 173 | 3 | 479 | 655 |
| No-equivalent | 11 | 181 | 514 | 706 |
| Imposter | 5 | 3 | 587 | 595 |
| Totals | 1,532 | 1,729 | 1,760 | 5,021 |

long-distance, dialed-up connections. The composition of the test set, shown in Figure 4, is 61-percent *yes* and *no*, 27-percent yes-equivalent and no-equivalent, and 12-percent imposters. The curves in Figure 3 show that it is possible to achieve an operating point below 1-percent FA and 5-percent FR on national data such as these.

Table 2 is a detailed confusion matrix for the March 1989 experiment using an operating point of approximately 1-percent FA and 5-percent FR. The different rows of the table correspond to what the caller actually said (*yes*, *no*, yes-equivalent, no-equivalent, or imposter) and constitute the input to the recognizer. The columns correspond to the three possible recognizer actions: accept as *yes*, accept as *no*, and reject. Each cell in the table shows the number of tokens out of 5,021 in which the specified input produced the specified recognizer output. For example, the first row of the table shows that out of 1,439 *yes* tokens, the recognizer classified 1,336 as *yes* and 20 as *no* and rejected the remaining 83. The fifth row of the table shows that of the 595 times that callers said words outside the recognizer's vocabulary (imposters), the recognizer mistakenly recognized a *yes* five times and a *no* three times and correctly rejected the input 587 times.

Results have been similar when real customers have used a speech recognition system to place collect and third-number-billed calls in the Grand Rapids, Michigan, LATA: The FA rate is less than 1 percent and the FR rate is less than 5 percent. An operating point at which the FA rate is lower than the FR rate was chosen because the perceived cost of a false acceptance is substantially higher than that of a false rejection. When a rejection occurs, the caller is given a second chance to respond; if a second rejection occurs, the caller is transferred to an operator. On the other hand, a false acceptance triggers an undesired action: billing for an unwanted call or termination of a desired call.

On May 5, 1989, at 6:59 a.m. in Grand Rapids, the first public-customer collect call was automated by means of speech recognition. The call was from "Glenn" and worked perfectly. The system was put into full service on May 15, 1989.

VSN systems have been deployed in 36 sites in the Ameritech and NYNEX regions.* In an Ameritech study of 2,608 tokens from public users, spoken in response to the first prompt for input, 24 tokens (0.92 percent) were false acceptances, while 47 tokens (1.8 percent) were false rejections.[11] The remaining 2,537 tokens were handled correctly. A bilingual version of the VSN is scheduled for introduction in Bell Canada this year.

So far, customer satisfaction with AABS has been high. We feel that this is due to two factors: the attention to detail that went into the design of the voice dialogue and the excellent rejection characteristics of the recognizer when confronted with imposters. We gained invaluable experience through the 1988 Bell Canada 976 Directory concept trial, which guided our thinking and forced us to focus on these two key issues.

Potential future applications of speech recognition in the telephone network include voice entry of telephone calling card numbers, catalog shopping order entry, voice control of voice and text message systems, automation of additional operator services, control of subscriber calling features, voice dialing for mobile telephones, airline flight status information, frequent-flyer account status, real-time financial information retrieval, automated

*Ameritech is the parent company of Michigan Bell, Ohio Bell, Indiana Bell, Illinois Bell, and Wisconsin Bell. NYNEX is the parent company of New York Telephone and New England Telephone.

switchboards, bank by phone, student course registration, and talking yellow pages. ■

## Acknowledgments

The INRS-Télécommunications research on very large vocabulary speech recognition was supported by the Natural Sciences and Engineering Research Council of Canada.

The author wishes to thank the following colleagues, who made significant technical contributions to the work described in this article: Greg Bielby, Doug Sharp, C.C. Chu, Andrew McGregor, Paul Boucher, Vishwa Gupta, Richard Jankowski, Georges Mony, Pierre Gendron, Ed Dermardiros, David Sloan, Rafi Rabipour, and Paul Mermelstein.

## References

1. M. Lennig and P. Mermelstein, "First Public Trial of a Speech-Recognition-Based 976 Directory," *Proc. Speech Tech '88*, Media Dimensions, New York, Apr. 26-28, 1988, pp. 291-292.

2. L. Deng, M. Lennig, and P. Mermelstein, "Modeling Microsegments of Stop Consonants in a Hidden Markov Model-Based Word Recognizer," *J. Acoustical Society of America*, Vol. 87, No. 6, June 1990, pp. 2,738-2,747.

3. L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, Mar. 1983, pp. 179-190.

4. K.-F. Lee, H.-W. Hon, and R. Reddy, "An Overview of the Sphinx Speech Recognition System," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-38, No. 1, Jan. 1990, pp. 35-45.

5. M. Lennig and M.L. Hanford, "Automating Operator Services with Speech Recognition," in *Voice Processing: Cashing In on the Telephone Network* (Proc. Probe Research Voice Processing Conf.), Probe Research, Inc., Morristown, N.J., 1988.

6. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, 1980, pp. 357-365.

7. M. Lennig, P. Mermelstein, and V.N. Gupta, "Speech Recognition," Canadian Patent No. 1 232 686, issued Feb. 9, 1988, Ottawa, Canada.

8. L.F. Lamel et al., "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 4, Aug. 1981, pp. 777-785.

9. M.J. Hunt, M. Lennig, and P. Mermelstein, "Use of Dynamic Programming in a Syllable-Based Continuous Speech Recognition System," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and J. Kruskal, eds., Addison-Wesley, New York, 1983, pp. 163-187.

10. V.N. Gupta, M. Lennig, and P. Mermelstein, "Decision Rules for Speaker-Independent Isolated Word Recognition," *Proc. 1984 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, IEEE, Piscataway, N.J., 1984, pp. 9.2.1-9.2.4.

11. R.W. Bossemeyer, E.C. Schwab, and B.A. Larson, "Automated Alternate Billing Services at Ameritech," *J. American Voice I/O Society*, Vol. 7, Mar. 1990, p. 50.

**Matthew Lennig** is manager of interactive services for Bell-Northern Research, with research and development responsibilities in the speech and image processing areas. Previously, as manager of interactive voice systems, he was responsible for development of the speech technology component of Northern Telecom's automated alternate billing service, described in this article. Earlier he was manager of speech systems, responsible for development of Bell Canada's speech-recognition-based 976 Directory and for algorithmic research in speech recognition, speaker verification, and speech synthesis.

Since 1981 Lennig has been a visiting professor at the University of Quebec's INRS-Télécommunications, where, from 1986 to 1989, he headed an NSERC-funded research project on very large (86,000-word) vocabulary speech recognition.

Lennig graduated summa cum laude from Princeton University in 1974 in an independent concentration combining mathematics, linguistics, computer science, and electrical engineering. He received a PhD in linguistics in 1978 from the University of Pennsylvania, which he attended as a National Science Foundation fellow, and an MEng in electrical engineering from McGill University in 1984. He is a member of the IEEE Computer Society.

The author can be contacted at Bell-Northern Research and INRS-Télécommunications, 3 Place du Commerce, Nuns' Island, Montreal, Quebec, Canada H3E 1H6.