ELSEVIER

Speech Communication 23 (1997) 31–39

SPEECH COMMUNICATION

The way we were: Speech technology, platforms and applications in the 'Old' AT&T

Robert J. Perdue [1]

Lucent Technologies, Bell Laboratories, Room 1B-386, 6200 E. Broad St., Columbus, OH 15303, USA

Received 25 February 1997; revised 30 April 1997

Abstract

The last several years have been an exciting time at AT&T in the field of advanced speech applications for telecommunications: technical progress and platform/processor advances have enabled the identification, development and testing of a range of new services. During this period, prior to the divestiture of AT&T of Lucent Technologies and NCR, AT&T brought together, under a single corporate 'roof', a research laboratory committed to advancing speech technology, business organizations building platforms to leverage this technology for telecommunications applications, and yet other business organizations with responsibility for deploying speech-enabled services to facilitate the use and reduce the cost of telecommunications services for both consumers and businesses. While this period of our corporate history has drawn to a close, we can look back to provide an overview of how technical progress, platform advances and network services needs and opportunities interacted to make speech technology an everyday experience for millions of people – and some of the lessons we learned along the way. © 1997 Elsevier Science B.V.

Résumé

Les dernières années ont été très intéressantes chez AT&T dans le domaine des applications avancées du vocal en télécommunications: des progrès techniques et des avancés au niveau des processeurs/plateformes ont permis l'identification, le développement et le test d'un ensemble de nouveaux services. Pendant cette période, et avant la séparation d'AT&T, de Lucent Technologies et NCR, AT&T rassemblait, sous un même 'toit', un laboratoire de recherche engagé sur les technologies vocales avancées, des organisations d'affaires construisant des plateformes pour pousser ces technologies dans des applications en télécommunications, et d'autres organisations d'affaires ayant la responsabilité du déploiement de services à technologies vocales pour faciliter l'usage et réduire le coût des services de télécommunications pour les abonnés et les professionnels. Maintenant que cette période de notre histoire commune prend fin, nous pouvons regarder en arrière pour dresser une vue d'ensemble sur comment les progrès techniques, les avancées au niveau des plateformes et les besoins et occasions au niveau des services centralisés (à travers un réseau téléphonique) ont interagi pour que la technologie vocale devienne une expérience quotidienne pour des millions de gens – et pour décrire quelques leçons que nous avons appris au cours de ce chemin. © 1997 Elsevier Science B.V.

Keywords: Connected digit recognition; Final state grammars; Interactive voice response; Speaker dependent recognition; Speaker independent recognition; Speaker verification; Speech recognition; Speech technology; Telephone operator automation; Voice dialing

[1] E-mail: rjperdue@lucent.com.

0167-6393/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved.
PII S0167-6393(97)00035-6

1. Introduction

Speech communication is the foundation of civilization. The use of tools built a technological society. The ability to converse with our tools has been little noted but will a have far-reaching impact. We have crossed a great technological threshold but it is anticlimactic. Science fiction has presaged it for decades and people communicate by speech so effortlessly that most do not think of hearing and understanding as a difficult problem. In the future, customers will want to interact with their machines as extensions of themselves in every possible way. Although now in an embryonic stage, the technology for humans to direct machines by speaking to them will be profoundly powerful and valuable.

This paper describes several successful services offered by AT&T based on speech recognition. A key element of developing successful services, that will not be covered, is managing the customer's expectations. When non-speech-research people hear the term *automatic speech recognition* their expectations are set by what they are familiar with – human speech recognition – against which the present state of the art pales in comparison. Achieving a successful service requires acquainting the buyer of the service and users of the service with the constraints of the technology so that they are not disappointed that it is not a complete human replacement.

2. Methodology

During our development of numerous successful speech technology applications both in the network and in commercial applications, the need for both technical and market trials to get the highest level of applications and technology performance has become increasingly evident. Whether the speech technology (be it speech recognition, speech synthesis, or speaker verification) can handle the task at hand to the customer's satisfaction cannot be determined in the laboratory. Frequently, laboratory data do not properly reflect how people will really react to the technology in service; particularly if the prompts are changed between the time of the original speech data collection for the lab and the usage in the field trial. Even in those cases where the lab speech data do represent the field situation, it is sometimes difficult to reproduce field results in the lab. A method of iterative trial and refinement is described below, which has been found to produce successful services reliably.

3. VRCP extensions

The basic automation of the telephone operator's task in the AT&T network by the project known as Voice Recognition Call Processing (VRCP) has been previously documented (Longenbaker et al., 1994). The initial system, first deployed in 1992, featured, ostensibly, a recognition vocabulary of {*collect, calling card, third number, person* and *operator*}. The vocabulary was actually somewhat larger, to allow for some common vocabulary synonyms, like *collect call, third party* and *person to person*, but it also employed Finite State Grammars. It was always naive to expect callers to only speak back a single element of the vocabulary. Callers are wont to say things like: ''I'd like to place a *person-to-person, collect call*, please, *operator*.'' In this example, the caller requested *person to person* treatment as well as *collect call* treatment, besides adding the superfluous, but polite, ''please *operator*''. Present operator practices specify how a call is to be treated if a caller says both *person-to person* and *collect*. The VRCP recognizer is required to process several such multiple vocabulary word call request variants in specified manners that closely match what operators do. Finite State Grammars that accommodate Filler Words provide the facility to allow for the specification of the disposition of caller utterances that contain multiple vocabulary words. This system, as initially deployed, has been tuned to give the performance depicted in Fig. 1.

Since 1992 the service has been expanded to recognize spoken credit card numbers and spoken

Correct Recognition	>96%
False Rejection	1%
Confusion (Substitution)	< 1%
False Acceptance	1% - 2%

Fig. 1. Recognition performed in VRCP.

telephone numbers for third number billing. Variable length connected digit string recognition that would be used to distinguish between different length credit card numbers from different credit card issuers has been trialed but not deployed. In addition, 00-traffic (the callers dialed 00 only, with no follow on telephone number) is now automated through the VRCP platform. Even though, 00- is the traditional number to dial when callers are most unsure or tentative about what they want or how to go about getting it (''If you need help, just dial 0 for an operator''), we have had good success automating this task. Providing billing credits for previous calls and information about area codes are the additional functions that have been automated. At the 00- initial prompt the caller is requested to speak one of the vocabulary words {*espanol, billing credit, information* or *operator*}. Based on the spoken vocabulary recognized, a series of menu subtrees follows, in which the caller speaks single digits to select elements from the menus to further specify the caller's request. The recognition models were initially built from speech data collected from the live operator/caller interaction, before the automated system was enabled. The percentage accuracy was between 80% and 90%, largely, due to the mismatch in conditions between the actual service and the environment in which the speech samples used for models were collected. In a subsequent iterative step, we tuned the weights in the grammar (the relative weight attached to the various elements of the vocabulary, including the filler models that soak up extraneous audio) and added speech data collected live from the actual service to raise overall accuracy over 90%. The speech recognition system in VRCP now handles about 3 million calls a day. Exact dollar costs saved by this system are not available to the public, but one can guess at order of magnitude savings. If the automated speech recognition system saves between 1 cent and 10 cents on each call, then, at an estimated average call volume of 3 million calls a day, savings are around $10 million to $100 million per year.

4. AT&T Voice Line™

The AT&T Voice Line™ trial (AVL), and later service, offered sophisticated speech technology for subscribers of the service. It was initially conceived to trial, for selected customers, custom calling features, such as, Voice Dialing from a Personal Calling List, secure account access though account numbers and speaker verification, Voice digit dialing, Voice Commands to control administration of personal account, and recognition of some globally known keywords to access features generally available (Fig. 2). As a consequence it incorporated several forms of speech recognition and speaker verification.

The AT&T Voice Line trial had several phases. During the trials, about 400 'customers' used the system as a functioning service. The 'customers' were selected to match a demographic profile that corresponded to the customer set to whom the service would eventually be offered. Mostly, the users were non-technical business people that would benefit from, and be willing to pay for, this type of service. In addition, the later phases of the technical trial and the initial offering of the real service to thousands of customers, overlapped in time. The different phases had different features to be tested and were closely controlled to gauge performance. This technique was useful as a continual quality refinement tool. Even after the service was offered, technical people were still on the project, although working on the trials, and were interested and available to observe, analyze and tune the speech performance in the real service.

A methodology for refining the speech technology development for applications that, so far, has proven to lead to reliable and technically successful services, was employed for the Voice Dialing trial. Both during the trial phases, and frequently during the service offering, it was essential to refine the prompts, based on caller responses, to elicit correct responses. From there, features and capabilities were added to

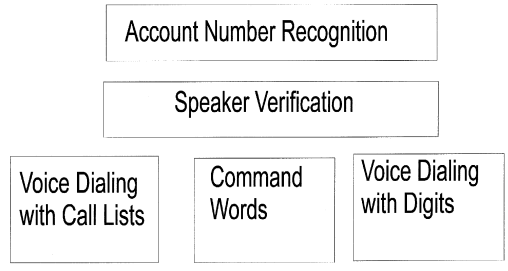

Fig. 2. Speech features in AT&T Voice Line™.

the Voice Dialing system. The procedure we follow is to build a prototype system that can handle the transactions for the trial population. As the trial proceeds, the speech data of each participant's responses to the system are collected. In the lab, technicians listen to the spoken caller responses, and transcribe what was said in order to label the samples. Based upon this labeled speech database, the system accuracy can be calibrated by comparing the recognizer's output with the technician's label. Error modes can be assessed. The first priority is to look for egregious errors in the transaction or failures of the technology. During the course of the trial, improvements are made in the lab to the recognizer or the speech models in the recognizer. It is essential to have methods in the lab that reliably predict field performance, so that experimental changes to the recognizer or the models can be gauged as improvements within the confines of the laboratory. Methods that subject the experimental recognizer under test to almost the exact field environment yield results that most reliably predict field performance. Further caution is needed to assure that a particular corpus of speech data is not overworked, that is, that the corpus is not used so much to risk that the recognizer is improved by tuning to the speech corpus. The risk exists that the recognizer may be adjusted to work spectacularly well for that particular body of speech samples but may show no improvement or, actually, suffer performance degradation when presented with the next batch of speech samples. As the trial progresses, this iterative process of collecting and labeling speech samples from the operating service, then assessing and improving performance, continues.

The Voice Dialing system allowed subscribers to enter the system and then, for instance, be able to make voice dialed calls from their personal voice label list – simply by speaking a label. To access their calling lists and other services, users first identified themselves. That access needed to be secure. There were to be millions of other accounts, but access had to be quick. Yet, both false acceptance of impostors and false rejection of true subscribers had to be rare.

To accomplish these goals for secure access, we employed speaker verification (SV) technology (Rosenberg et al., 1994) in combination with the spoken entry of an account number. This SV technology is related to speech recognition in that the underlying HMM algorithms are similar. Although Speaker Verification technology is not as extensively studied as speech recognition, we felt that the performance would be acceptable to meet the needs of the Voice Dialing system. Moreover, we pursued efforts to insure fast system response. For example, during enrolment a subscriber spoke the verification phrases three times. Voice models calculated from those utterances were compared to many other voice models and were grouped with those that were most similar. Subsequently, whenever a caller's speech was verified, the test utterance was compared only within that subscriber's group of similar voices. Differentially comparing the test utterance with the subscriber's speech and the subscriber's group avoids exhaustive real-time comparisons against all other stored subscribers' speech.

Generally, during the verification phase in a speaker verification system, a user performs two actions. The user first makes an identity claim and second says an identification phrase. The user's speech pattern is compared against the person's speech pattern associated with the claim that was deposited during enrolment. In contrast, we chose to develop an efficient, non-obtrusive method for security during normal usage that was simpler than the two-step approach.

Our solution in the Voice Dialing system is to have the user make an identity claim by saying the account number. From the user's perspective the trial used a single-step process shown in Fig. 3. The system performs connected digit speech recognition on the spoken phrase to identify the account. It

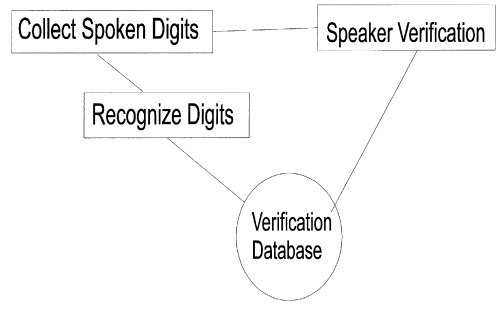

Fig. 3. The single step process for speaker verification using account number recognition.

fetched the speaker verification voice models for that account and then performed the verification task, comparing the spoken account number with the voice models associated with the account.

Two distinct trials were conducted to evaluate the speaker verification performance. The first trial was used by the trial population as they would in real service. They received instructions not to allow non-participants to try their account. Next an impostor trial was conducted, in which 100 males and females were given numerous other legitimate account numbers (gender-matched) and were told to try to break in.

Overall performance of the system in the trial was excellent. Once the connected digit account number has been successfully recognized, the speaker verification accuracy was high – first attempt callers were accepted 95% of the time while impostors were accepted less than 1%. When given a second attempt, legitimate callers were accepted 97% of the time while the impostor acceptance rate rose to about 1.5%. By varying the threshold of acceptance, the rates of impostor acceptance versus false rejection could be traded off against each other. Although the performance of SV exceeded our expectations, it did have vulnerabilities and has not been incorporated into the actual service. There were two main problems, both a result of the usage of the SV system and not due to the technology per se. First, the requirement to minimize the enrolment time, primarily to minimize user annoyance, limited the number of enrolment phrases to three. This user annoyance factor is a strong consideration. Subscribers did not want to say number strings three times, but seemed resigned to do so. They complained markedly more when enrolment required more than three repeats; yet more enrolment phrases are needed to ensure high accuracy. Our solution was that, subsequent to the brief enrolment period, in which the subscriber spoke the verification phrases only three times, the system continued to adapt, for a few calls, to the caller's voice until it reached a certain level of stability. If an impostor broke in during the adaptation phase then the system became very vulnerable. An impostor break-in during the adaptation phase is a possible, though not very probable, event unless the subscriber and impostor are in collusion. The second vulnerability of the system was due to some customer behavior that was not expected in service. By design, the Speaker Verification feature was never intended to be the main security barrier, but was to be an additional layer of security on top of existing account numbers. Subscribers know, and people in general know, not to give their account numbers to anyone else. They know that, in the typical systems they use in life, with only account numbers as security and, of course, no Speaker Verification, if they give their account number to an impostor that impostor can immediately break in to their account. The addition of Speaker Verification to a system is attractive because now the impostor has to both know the account number and sound like the subscriber. However, despite admonishments about security in the trial, subscribers would occasionally tell others their account numbers in order to show the efficacy of the Speaker Verification. Were this done during the adaptation period, break-in could occur and the reputation of the Speaker Verification would suffer.

As previously stated, speaker-independent Connected Digit Recognition was used for account number recognition even without the speaker verification step. Since most of the potential subscribers to the Voice Dialing system already have Calling Card account numbers, it was decided to use these numbers as the Voice Dialing account number. While that may be convenient to new customers, it does not necessarily enhance recognition performance. Connected digit accuracy degrades somewhat, particularly in noisy environments, when the string length is unknown. Previously assigned calling card numbers lack built-in check digits and have very little grammatical structure to the strings. Since the card numbers had been assigned at different times by different organizations, there are few rules of structure (*e.g., the first digit is always 3 or 5, or the fourth digit is always zero*) as there are for commercial credit cards. However, first and second choice recognizer candidates can both be checked against the database for acceptance, and overall account number recognition accuracy is greater than 90% on the first attempt.

An important feature of the system is the ability of subscribers to train a set of *voice labels*. Each label has an associated telephone number that is dialed when the subscriber pronounces the *voice label*. Generally, it was envisaged that the subscriber

would speak a *voice label* a few times, associate a telephone number with it and thereby enrol into the system. Many techniques exist to accomplish the task of speaker-dependent recognition for tasks, such as *voice dialing*. Some key requirements are that the system be easy for the subscriber to customize with his or her own vocabulary. If fact, if it is possible to give the subscribe different options for training the system, that is desirable. Many of the methods for providing speaker-dependent recognition require a large amount of storage per user. If there are to be many eventual users of the system then a large per user storage forces the design of a system with a great deal of, at the time, expensive storage. The costs of high capacity disc drives have dropped dramatically over the past few years, so the expensive of just raw memory is less of a problem. However, fetching and shuffling about large amounts of user data does put a burden on the data transport facilities of the system and, in the case of shared database servers, the input/output capabilities of the database machine. Of course, these requirements do not dictate use of speaker-dependent technology at all, only a speaker trained technology. We have developed a speaker-dependent technique that employs speaker independent phonemes as the basic alphabet of the speaker dependent system. When a subscriber enrols in the system by training the *voice labels*, each utterance is matched against the set of speaker independent phonemes to derive the *voice label* pattern. The resultant pattern is frequently a poor orthographic transcription of what the subscriber actually said but this is of little consequence. It is unnecessary for the pattern to be an exact transcription of the utterance. It is only necessary to be a distinctive 'audio signature' that can be discriminated from the other patterns in the subscribers *voice label list*. At that, it works well. It has a further slight advantage that patterns built of speaker-independent phonemes allow others to use the subscribers voice labels, an attribute not widely shared by other speaker-trained systems. The use of speaker-independent phoneme models yields the advantages that the *voice label list* can be built either by the subscriber speaking the *voice labels* a few times or by prior written submission of the names from which the phonemic transcription can be determined. Since the patterns that are stored that model the entries in the *voice label list* are relatively short sequences of phonemes, the storage requirements of the *voice label* system are small. Subscribers are advised to make their *voice labels* multi-syllabic to improve recognition and minimize inter-label confusion, but no actual enforcement occurs.

The trials incorporated a *closeness-checking* mechanism during voice label training to inform the subscriber that a newly entered voice label was too similar to an existing label. The speaker-trained recognizer *incorporates key word spotting* and *non-vocabulary rejection*, but extraneous speech is not a frequent occurrence, primarily because the subscribers are trained and are motivated by self-interest to have their transaction flow as quickly and efficiently as possible. In addition, prompt interrupt capability was provided for two main reasons. When the prompts are long, expert users want to interrupt the prompt to hurry their transaction along. Even when the prompts are short, expert users still have a tendency to anticipate the end of the prompt and start speaking slightly prior to the prompt completion. If prompt interrupt is not allowed and the user starts speaking before the prompt ends, only the fragment of speech after the end of the prompt gets passed to the recognizer. In addition, *key word spotting* becomes problematic. In cases where the subscriber enrols a name like *Smith's* and its associated home phone number and then enrols *Tom Smith* as an office number, for instance, it is very easy for an aggressive *key word spotting* recognizer to always ignore the *Tom* in *Tom Smith* and misrecognize it as the other label *Smith's*. Consequently, the key word spotting capability must be tempered.

Overall performance of the speaker-dependent voice label system under trial has been excellent – first attempt label recognition accuracies are in the 80–90% range. Furthermore, levels of performance and customer satisfaction will improve with experience. Subscribers will derive benefit from proper operation of the service and so will be motivated to have the system work well. They will be inclined to modify their behavior slightly if it is necessary for successful operation. Conversely, if the system really does not work for them, they will drop out of the service.

At the time, it was planned that the service would eventually offer voice commands and provide callers

with a set of *universal voice labels* that would be available on everyone's voice list. The same machinery we used for the speaker trained voice labels was also useable for voice commands and speaker independent voice labels. Trials were conducted testing the operation and acceptability of *voice commands* for administering *voice label lists*. Subscribers could elect to use speech to *add labels*, *delete labels*, *review labels*, etc., rather than employing DTMF menus. Since list administration would be an infrequent event once it was initially built, voice commands give the user a more intuitive method, or at least, a more easily remembered method for administering their account than DTMF.

5. Universal Card Services™

At AT&T's Universal Card Service any holder of a Universal Card can call to request financial information. Single digit speech recognition is used to traverse menu trees to locate the correct information source. Callers speak their 16 digit Universal Card number to the system to gain access to their accounts. Once an account number is recognized, the caller is asked to speak his or her five digit ZIP code as further confirmation of access privilege before information, like an account balance, is divulged. Callers speak their Universal Card numbers and Zip code numbers as they would to a human attendant – deliberately, smoothly, and, of course, with no required pauses between digits. The thousands of callers that call every day were never trained or informed how to speak to the system. They are allowed to just speak normally. The Lucent Technologies INTUITY™ CONVERSANT™ product speech recognizer that performs this connected digit string recognition is a Continuous Density Hidden Markov Model based recognizer. Finite state grammars are used to describe the length and form of the legal digit strings. Callers may *barge-in*, that is interrupt the prompt, so the recognizer has to be quite tolerant to extraneous speech or background sounds that can be presented to the recognizer. Excellent echo cancellers reduce the amount of prompt that is reflected back to the recognizer's input. However, even during long prompts, callers frequently wait for the prompt to end before speaking a vocabulary word, during that interval when the caller is waiting for the prompt to finish, large amounts of background noise and inadvertent caller noise can be presented to the recognizer's input. A good recognizer must reject that extraneous noise and not false trigger by either giving a false recognition or turning off the prompt. A discriminative training technique called Generalized Probabilistic Descent (GPD) (Huang et al., 1995; Mikkilineni and Webb, 1996) was used in model training to minimize string errors and errors caused by extraneous sounds. The metrics of this technique maximize the distance between vocabulary models, and filler models for silence and extraneous sounds, by focusing on minimizing the overall string error rate of the training sets. The speech data used in training were gathered from a previous trial with Universal Card Services as well as from the VRCP credit card entry trial. Our methodology of iterative refinement by collecting speech data in situ, improving the recognizer's performance based on those data, and then collecting more speech and field performance data to again refine the recognizer was used to get the string recognition rate above 90%. A final step of comparing both the first few string choices from the recognizer output against the database of legitimate card numbers raises the overall card number acceptance rate even higher.

At present, other technologies are being explored for possible incorporation into Universal Card Service transactions. When customers apply for a credit card from UCS, one of the items on the application form is for the customer's mother's maiden name. As a security measure, to control access to some account information, attendants request that the caller say her mother's maiden name so that it may be confirmed against the account database for that caller. Only attendants can confirm the mother's maiden name, now, so a lengthy transaction must involve a human when most of the transaction could be otherwise automated. Subject to the continued acceptability of mother's maiden name as a security technique, it would be economical to automate this confirmation step in order to mechanize the complete transaction. A variant of the Lucent Technologies INTUITY™ CONVERSANT™ FlexWord recognizer is being trialed in this regard. A key element of this recognizer variant is the ability to perform excellent

out-of-vocabulary rejection. After all, the vocabulary only has one word – the correct maiden name – and anything else that is said must be rejected. In operation, during the call, the text for the maiden name is fetched from the account database, translated into a phonetic or subword spelling for the recognizer and then downloaded into the special FlexWord recognizer. Variations in name pronunciations are large, but the system does achieve a high enough level of automation to be attractive. A second application of technology is to add a ''How may I help you?'' prompt at the beginning of the transaction rather than a menu prompt. This *natural language* approach to the recognition problem is a further step down the path to more conversational dialogs with machines. This system is not in trial yet.

6. Other network applications of speech recognition

We have experience with other telephone network services that incorporate speech recognition and, like VRCP, the Voice Dialing system, and the transactions at Universal Card Services these services also require the Speech Recognition (SR) technology to be very robust to extraneous speech, non-responsive inputs, and varying levels of background noise.

The AT&T True Ties™ service (personal 800 numbers) uses speech recognition for entry of Personal Identification Numbers (PIN). For 800 numbers the caller receives a free call. All calls are paid for by the called party. When individuals possess personal 800 numbers, the financial consequence of receiving a large volume of wrong numbers can be devastating. This can occur when the personal 800 number is only one digit different from a major commercial 800 business number. Therefore, owners of personal 800 numbers are assigned a PIN. No one can complete a call to that number without entering that PIN. The subscriber, of course, distributes that number to all parties that she expects to call her 800 number. Whenever people call a personal 800 number they are instructed to say or enter the PIN for that number before the call is completed and charged.

Another service that uses SR is an AT&T directory assistance service. In this application, when callers call AT&T and inquire about telephone numbers, the requested telephone number is stored and played from a Local Exchange Carrier directory assistance database to which AT&T has no data connectivity. Therefore, when the local directory assistance service plays the telephone number to the caller, speech recognition in the AT&T system recognizes the spoken telephone number and can then offer to place the call if the caller wishes. This task is best accomplished using speaker-dependent trained speech recognition technology. Recordings of the telephone number announcements used in all the announcement systems around the country are sampled and built into recognition models. The number of voices employed and the range of vocabulary pronounced is limited. However, the requirements on recognizer accuracy are extremely high. In service, telephone number digit string recognition rates are typically above 95%. The recognition accuracies are so high that, whenever the recognition algorithm scores below a high threshold level of confidence, a process is triggered to begin sampling recorded number announcements in that area. The accuracies are so high that cases of missed recognition are almost always indicative of the signing on of new or different recorded announcements into the system.

7. Implementation

All of the services described in this paper were implemented using Lucent Technologies' INTUITY™ CONVERSANT™ product, a voice processing system based on many open standards. This system can stand alone or be used as a building block to assemble large-scale services that connect to various network switches. Within the system, a set of signal processing boards is used to perform the real-time recognition in a multi-channel environment. Similar configurations are employed in both network and commercial environments around the globe.

8. Future directions

Speech communication with machines is profound. It allows many practical tasks both to lower costs of existing services and to provide new ser-

vices previously uneconomical because they required a person.

The ultimate goal is to provide a machine replacement that is practically indistinguishable from a person, regardless of the application. We are a long way from that goal. However, we will be able to approach the goal in a few well defined and well designed applications. The speech technology we have described in this paper is language-independent. Consequently, there should be good growth in the deployment of network applications incorporating speech technology around the world.

More conversational, easy to use, recognition systems will be available over the next few years but will require a great deal of application development work. Continued work will be needed to structure the user interface of the applications according to sound human factors principles. Additional effort will be needed to gather customer speech to build a good model of how callers interact with the system, in order to train the system to accommodate those interactions properly. Other work will include the construction of sentence grammars for expected responses, methods for determining when sufficient correct information has been gathered, and methods for eliciting good data to fill those areas where information is insufficient. A great deal of study of written language and how to model it has been done to allow for the construction of dictation machines that are quite impressive. However, people do not converse with each other using grammatically correct prose. In fact, they frequently do not even use complete sentences. A greater understanding of spoken, conversational dialog is needed to build a machine that can carry on a good conversation.

Other important technology trends that will influence SR-based system design include the development of personalized user interfaces with intelligent agent-like functionality; the increased networking and integration between customer-premises-based systems and large-scale telephone networks; merger of functionality between traditional messaging and interactive voice response systems; and the addition of other media to the user interface.

References

Huang, B.H., Perdue, R.J., Thomson, D.L., 1995. Deployable automatic speech recognition systems: Advances and challenges. AT&T Techn. J. 74 (2), 45–56.

Longenbaker, W.E., Perdue, R.J., Salchenberger, S.M., 1994. Automation of operator services: A successful application of speech recognition technology. In: Proc. 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA 94), September 1994, pp. 161–164.

Mikkilineni, P., Webb, J.J., 1996. Discriminative training of a connected digit recognizer with fixed filler models and its application to telephone network service systems. In: Proc. 1996 IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, 30 September–1 October.

Rosenberg, A.E., Lee, C.H., Soong, F.K., 1994. Cepstral cohort normalized techniques for HMM based speaker verification. In: Proc. Internat. Conf. on Speech and Language Processing, September 1994.