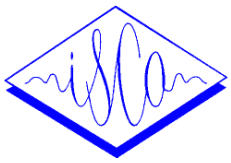

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Bellcore Efforts in Applying Speech Technology to Telephone Network Services

G. Velius, C. Kamm, M. J. Altom, T. C. Feustel, M. J. Macchi, M. F. Spiegel

Bellcore, Morristown, NJ, USA

Introduction

Recent speech technology research at Bellcore has contributed to several telephone network applications that are now either being deployed or tested. These applications include *Automated Alternate Billing Services* (AABS) for collect and third party calls using automatic speech recognition; reducing operator connect time in *Directory Assistance* (DA) via speech compression; an *Automated Customer Name and Address* (ACNA) service using text-to-speech synthesis; and remote security services such as credit card validation and home-incarceration checks using Speaker Identity Verification (SIV). In this paper, we describe each of these applications and the speech technology and human-interface issues critical for successful deployment.

Automated Alternate Billing Services - AABS

A telecommunications equipment manufacturer has provided speaker-independent Automatic Speech Recognition (ASR) technology for introduction into the telephone network to automate the acceptance or rejection of charges for collect and third number billed calls, known as AABS. In a collect call, the system asks for the caller's name and records it. The recording is then used in announcing the collect call to the called party. The called party is then asked if he/she will pay for the call, and the response is automatically recognized as affirmative, negative, or indeterminate. If the response is recognized as affirmative, the connection between the calling and called parties is completed. If the response is recognized as negative, the connection is not made, and the calling party hears an explanatory announcement. If the system cannot determine the response with high certainty, an operator intervenes to complete the transaction.

The AABS application nominally requires a vocabulary of only two words ("yes" and "no"), a capability well within the current state of the art in speaker-independent isolated-word recognition. However, our observations of customer interactions with the system show that the application is not quite so simple. In a pilot study, customers were prompted to give their responses "after the tone", yet nearly 50% spoke before the tone. The pilot study also showed that, even with a prompt specifically asking for just the word "yes" or just the word "no", 20% of the customers still say something other than an isolated-word response of either "yes" or "no". Another difficulty is that some customers, perhaps confused, will repeat the word "hello", which may be misrecognized as "no". Furthermore, on about 10-15% of the AABS calls that are answered, the first response is from someone who cannot authorize alternate billing to that number or an answering machine. Still, recent AABS trials described by Bossemeyer *et al.* (1990) [1] have shown that more than 75% of users successfully complete the automated interaction.

Among the lessons we have drawn from our experience with AABS thus far are that 1) a capability for recognizing responses spoken during the prompt is highly desirable, if not essential; 2) even with carefully worded prompts, it is difficult to control the dialog completely; and 3) restricting the ASR vocabulary to the two words "yes" and "no" may not provide a sufficient vocabulary for successful deployment of the AABS application. The AABS service is a demonstration that user-interface issues are just as critical to the successful use of speech technology in the telephone network as the performance of the underlying technology itself.

Speech Technology in Directory Assistance

Bell Operating Companies (BOCs) handle approximately 6.2 billion *Directory Assistance* (DA) calls each year. Campbell & Velius (1989) [2] have estimated that a decrease in operators' average work time of one second per call could free over 1000 DA operators to provide other types of services. Today, when a customer dials directory assistance and an operator is not immediately available, the call waits in a queue. While waiting, the caller hears audible ringing until the call is connected to an operator. When the call is connected to an operator, an announcement is played to the customer, the customer

speaks his/her request, and then the customer and the operator interact. Operator work time can be reduced in at least two ways. First, the operator can be brought into the connection after the announcement has been given to the customer. Second, the customer's speech can be processed to reduce the amount of time it takes to hear it. Time compression of the customer's request is accomplished by removing unnecessary silent intervals and parts of the speech signal that are redundant. We gathered durational statistics from actual DA calls and found that the average initial customer request of 4.1 seconds was reduced by about one second after applying our time-compression algorithm. A pilot study was done to determine whether operator keying time was actually reduced with the time-compressed customer requests. The average reduction in keying time was 0.6 seconds. Time compression resulted in poorer keying accuracy on 6 of the 196 customer requests tested in this study.

Typically, in this new DA scenario, a customer is first connected to a speech processing system and at some later time, after an operator has become available and has listened to the processed speech, the customer and operator are connected. This switchover to an operator may happen as the customer is still giving his/her initial request. In this case the customer's perception is that the operator has been on-line from the beginning of the call. If an operator does not come on-line until well after the customer has stopped speaking, the system asks the customer to please wait. Today's average customer delays can be maintained in the new DA scenario, but the customer delay moves from the beginning of the call (i.e. audible ringing) to the middle of the call. A field trial is currently under development to ascertain the long-term effects of compressed speech on operators, customer acceptance of the new service scenario, and operator work time savings.

In addition to the immediate benefits, the architectural changes introduced by this technology will provide a platform for further automation of DA with speaker-independent speech recognition. For example, customers could be prompted to speak the city name of interest, and if an ASR system could confidently determine what was spoken, the customer could be prompted further. If the ASR system is uncertain of what was said, an operator could hear the recorded city name and make the decision. With the decision parameters of the ASR system set so that decision errors rarely occur, an imperfect ASR technology can be productive in this scenario. Further improvements in ASR technology would then translate directly into a higher percentage of automation of the DA task, and greater savings in terms of operator work time.

Speech Synthesis

Current speech synthesis technology offers capabilities that can help expand the range of information available to telephone users. Many telephone network applications involve names and addresses. Examples include: *Automated Customer Name and Address (ACNA)*, where users input a telephone number and hear the directory listing for that phone number; *Who's Calling*, where the synthesized name of the caller is provided; and *Customized Intercept*, where a business's name is announced ("XYZ Corp has moved. Our new number is ..."). Our work on a demisyllable-based system, the *Orator*™ synthesizer, is aimed at these kinds of telecommunication applications. These services present special challenges to current speech synthesis technology, because users are generally inexperienced with speech synthesis, and because names of people, places, and businesses can be difficult to pronounce.

Commercial synthesizers were developed originally for pronouncing *words* according to rules of English pronunciation. In a test of the *name* pronunciation accuracy of four commercial English synthesizers, Spiegel (1985) [6] showed that the best synthesizers mispronounced 25% of the 2000 most common names in America. These high error rates for name pronunciation (compared to error rates for word pronunciation) demonstrated that the rules used to pronounce names are different from those used to pronounce English words.

Name pronunciation is difficult for several reasons. First, there is a great variety of names in America - over 1.5 million different surnames (family names) , plus many first names (given names), place names, and business names. In addition, the fact that names derive from dozens of source languages makes the writing of rules for name pronunciation quite difficult. Finally, there are problems associated with individualized pronunciations and varying degrees of Anglicization. We compared the *Orator* system to a high-end, widely-used commercial synthesizer using a a frequency-weighted sample of 1500 surnames, first names, and street names. Davoust and Spiegel (1990) [3] reported that the *Orator* system was judged to produce 5.2% mild errors and 1.1% severe errors, while the commercial synthesizer produced 10.8% mild errors and 4.1% severe errors.

One key component to accurate name pronunciation is the correct analysis of compound names. If a synthesizer does not decompose compound names

Orator is a trademark of Bellcore.

(e.g. "wineberger") into component morphs ("wine" and "berger," etc.), severe mispronunciations can result. One module of the *Orator* system analyzes each word in the input for morphs that are common to names. We currently recognize over 1,000 morphs, finding at least one morph in about 20% of all names.

Determining a sensible pronunciation of the information in business listings is a difficult task, because business listings contain a difficult-to-predict mixture of names and words. In fact, business listings may be the most widely-used database with such a broad coverage of both names and words (2/3 of the items are names). Business listings (as well as general computer text) also contain many acronyms, some of which are usually *pronounced* and others usually *spelled out*. The *Orator* synthesizer has a set of rules that determine whether an acronym is pronounceable or not. Short acronyms (two to perhaps four letters) are generally not pronounceable. For longer acronyms, rules attempt to break up the letter strings into syllables; if a string cannot be syllabified, the acronym is declared unpronounceable. Because no rule-based system can always be correct, the *Orator* synthesizer can optionally provide both spelling and pronunciation for intermediate-length acronyms.

In addition to high name pronunciation accuracy, these applications need high segmental intelligibility, because names and addresses contain little contextual information to aid the listener's comprehension of the words. Phoneme-based speech synthesizers often do not properly model the phoneme transitions and coarticulations present in natural continuous speech. The *Orator* synthesizer, however, is not based on phonemes, but rather is based on a demisyllable inventory that contains most of these transitions. Spiegel *et al.* (1988) tested the telephone intelligibility of the *Orator* synthesizer and that of the commercial synthesizer described previously, using a comprehensive monosyllabic corpus (including consonant clusters). This comparison showed that the *Orator* synthesizer has higher segmental intelligibility than the commercial synthesizer and scores higher in preference tests.[1]

Currently, several of the Bell Operating Companies are planning ACNA trials and/or services using the *Orator* system.

Speaker Identity Verification

Speaker Identity Verification (SIV) is a biometric security technique wherein an individual's identity is ascertained based on the unique way that he/she speaks. In a typical SIV scenario, a talker makes an identity claim by some means other than speech. Reference speech data from the talker is compared against a new sample of speech data from the same word or phrase solicited on-line. Based on the dissimilarity between the reference template of known origin and the newly solicited sample, a decision is made to validate the identity claim. If the dissimilarity is higher than a predetermined criterion, the identity claim is rejected; otherwise, the claim is accepted. Since this technology is an attractive and relatively unobtrusive means of providing security over telephones, several interesting applications are being investigated.

One of the Bell Client Companies is experimenting with our SIV system to understand its limitations with respect to a home incarceration application where a centralized SIV service would call, at random intervals, someone who is under court order to remain at home. This application is nearly ideal in three aspects: 1) the same telephone equipment would be used repeatedly, reducing or eliminating the network variability problem, 2) the user would have to interact with the system frequently (several times a day), thus making it easier to track changes in voice characteristics over time, and 3) the user presumably has substantial motivation to be cooperative with the system.

Another application being modeled in our laboratory involves using SIV to secure smartcard[2] access to payphones. The speaker's speech data are stored on the card itself so that the card is usable only by the owner. If the user's voice characteristics do not match those stored on the card, use of the phone is limited and access to information stored on the smartcard is restricted.

Over the past several years, we have conducted several studies aimed at increasing the accuracy of existing verification algorithms. This work includes extensive parametric analysis of our LPC-Cepstrum-based algorithm reported by Velius (1989) [8], as well as more limited studies on the relationship between the number of utterance comparisons and algorithm performance. In general, we have found that accuracy increases monotonically with the product of the number of stored reference templates and the number of solicited test utterances, at least up to a total of nine comparisons.

1. Since both synthesizers scored much lower than natural speech, we must conclude that for difficult applications (such as these), speech synthesis still needs to improve.

2. A smartcard is a credit card with a built in microcomputer.

In addition to our algorithmic work, we have done perceptual studies in an attempt to gain a better understanding of identity-bearing characteristics of the speech signal. Feustel *et al.* (1988) [4] found that human listeners could accurately distinguish the identity of speakers based on utterance comparisons that could not be distinguished by our original algorithm. Furthermore, it appeared that the information the listeners were using was contained in the part of the speech signal that was not adequately modeled by the algorithm. In a second experiment reported by Feustel and Velius (1987) [5], speakers' voices were monitored over a period of several months. Over that time, voice changes contributed to a near doubling of SIV error rates. This deleterious effect of time, however, could be ameliorated with simple template management techniques, including updating and storage of multiple templates.

Our research has resulted in an experimental verifier offering accuracies in excess of 96% for single, monosyllabic words at telephone bandwidths. Higher accuracies (>98%) are possible with longer or repeated utterances.

Summary

In summary, we have described four application areas of speech technology that are currently being used or considered for use in the U.S. public telephone network. These applications demonstrate that current speech technology, although somewhat limited, can be useful and usable, provided that other aspects of the application (e.g., the human-computer interaction) are carefully designed and structured. Furthermore, the knowledge gained from these early field applications will certainly facilitate further improvements in speech technology and lead to future deployment of speech technology in more complex services and applications.

References

[1] Bossemeyer, R.W., Jr., Schwab, E.C., & Larson, B.A., "Automated Alternate Billing Services at Ameritech", *Journal of the American Voice I/O Society Special RHC/RBOC Issue*, pages 47-53, Vol. 7, March, 1990.

[2] Campbell, J.B., & Velius, G., "Applying Speech Technologies to Directory Assistance," *Bellcore Digest,* Vol. 6, Issue 4, pages 1-8, July 1989.

[3] Davoust, J.E., & Spiegel, M.F., "How Well do State-of-the-Art Speech Synthesizers Pronounce Names?" *Proceedings of Speech Tech 90*, pages 347-352, 1990.

[4] Feustel, T.C., Logan, R.J., & Velius, G.A., "Human and Machine Performance on Speaker Identity Verification", *J. Acoust. Soc. Am.*, Supplement 1, Vol. 83, S55, 1988.

[5] Feustel, T.C., & Velius, G.A., "Long-Term Changes in Voice Characteristics: Implications for Speaker Identity Verification (SIV)", *J. Acoust. Soc. Am.*, Supplement 1, Vol. 82, S82, 1987.

[6] Spiegel, M.F., "Pronouncing Surnames Automatically", *Proceedings of American Voice I/O Society (AVIOS)*, Sept., 1985.

[7] Spiegel, M.F., Altom, M.J., Macchi, M.J., & Wallace, K.L., "Using a Monosyllabic Test Corpus to Evaluate the Intelligibility of Synthesized and Natural Speech", *Proceedings of American Voice I/O Society (AVIOS)*, Oct., 1988.

[8] Velius, G., "Variants of Cepstrum Based Speaker Identity Verification", *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Vol. 1, pages 583-586, 1988.