EXPERIMENTS WITH THE TANGORA 20,000 WORD SPEECH RECOGNIZER

A. Averbuch, L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza,
E. Epstein, D. Fraleigh, F. Jelinek, B. Lewis, R. Mercer, J. Moorhead, A. Nadas, D. Nahamoo,
M. Picheny, G. Shichman, P. Spinelli, D. Van Compernolle, H. Wilkens

Speech Recognition Group, Department of Computer Sciences
IBM Thomas J. Watson Research Center, P O Box 218, Yorktown Heights, NY 10598

ABSTRACT

The Speech Recognition Group at IBM Research in Yorktown Heights has developed a real-time, isolated-utterance speech recognizer for natural language based on the IBM Personal Computer AT and IBM Signal Processors. The system has recently been enhanced by expanding the vocabulary from 5,000 words to 20,000 words and by the addition of a speech workstation to support usability studies on document creation by voice. The system supports spelling and interactive personalization to augment the vocabularies. This paper describes the implementation, user interface, and comparative performance of the recognizer.

INTRODUCTION

Tangora represents a class of large-vocabulary, isolated-utterance speech recognizers based on an IBM Personal Computer AT and IBM Signal Processor subsystems [1]. These systems are named for Albert Tangora, listed by the Guinness Book of World Records as the fastest keyboard typist, who could sustain rates of 147 word per minute for one hour. The 5,000 word Tangora-5, presented at ICASSP in 1986, demonstrated the feasibility of a large-vocabulary PC-based recognizer, but offered only limited workstation function. This paper describes an enhanced Tangora configuration: the basic vocabulary has been expanded to 20,000 words, and a speech workstation has been developed to support usability studies on document creation by voice. The 20,000 word Tangora-20 represents the latest effort in a research project that began with work on large vocabulary continuous speech algorithms in 1972 [2],[3], and for the last six years has included work on isolated utterance recognition [4]-[6].

Tangora-5 uses two sets of Personal Instrument (PI) cards which fit within an IBM PC-AT. Tangora-20 uses four PIs and requires an additional AT-sized expansion chassis. The IBM PI Signal Processor subsystems, which are based on an IBM VLSI signal processor chip, provide additional memory, support analog I/O, and provide an interface to the IBM PC-AT bus [7]. Signal processing, vector quantization, and acoustic matching algorithms are programmed on the IBM PI Signal Processors.

IMPLEMENTATION

The recognition algorithms [6], [8]-[10] can be divided into the following modules: Acoustic Processing, Fast Acoustic Match, Language Model, Detailed Acoustic Match, and Hypothesis Search.

The Acoustic Processor is implemented in one of the PI subsystems. In demonstration versions of the Tangora, this subsystem and microphone preamplifier are contained within the same PC-AT as the rest of the decoder. Since the Acoustic Processor reduces the data rate from 30,000 to 100 bytes per second (20,000 12-bit samples to 100 single-byte acoustic labels), this component of the recognizer can be remote from the Tangora. By moving the Acoustic Processor subsystem to a separate PC-AT, multiple speech workstations can be supported by a single Tangora recognizer, connected through 2400 baud asynchronous dial-up modems. Only one user can be recognized at a time, but the more expensive resource can be shared over several users.

Language Model statistics are precomputed for a given vocabulary and are stored on a non-DOS disk partition, accessed by special software routines to meet real-time speed and data transfer requirements. The 5,000 word Language Model occupies about 15 Mbytes of disk, the 20,000 word Language Model requires about 18 Mbytes. During recognition, the data are read from disk and downloaded into one of the PI subsystems, which then computes the Language Model probabilities for the list of candidate words proposed by the Fast Acoustic Match.

In the Tangora-5, the Language Model calculations and both Acoustic Matches are performed in the same PI subsystem. In the Tangora-20 the Fast Acoustic Match is performed on three PI subsystems, two of which also handle the Detailed Acoustic Match, while the other handles the Language Model calculations. Faster recognition and greater accuracy can be obtained by adding more PI subsystems.

The Hypothesis Search is carried out on the PC-AT, running DOS, using 512 Kbytes of memory. The PC-AT also implements a rudimentary workstation with limited editing and display features on demonstration versions of the Tangora. In the remote workstation configuration, the PC-AT also handles the communication of labels and command strings from, and decoded word strings to, the remote workstation.

Tangora is a speaker-dependent recognizer. Each new user must go through an enrollment procedure to derive the prototype

CH2396-0/87/0000-0701 $1.00

vectors for the Acoustic Processor, and to train the Acoustic Model parameters. This involves reading a *training script* of 100 sentences, producing a speech sample of about 20 minutes. The current training script contains about 1200 words, only 700 of which are distinct, yet this is sufficient to train both the 5,000 word and the 20,000 word recognizers. This compares very favorably with word-based algorithms which might require multiple utterances of each of the words. The prototype vectors and Acoustic Model parameters for an individual speaker are relatively small (400 Kbytes), so that many speakers can be supported by a single Tangora using the PC-AT's hard disk.

All of the programs were first written in C, and developed and tested on VM/CMS, before being ported to the PC-AT. The computationally expensive routines were then migrated to assembler code running on the PI subsystems. All probability computations are carried out in the log domain on scaled 16-bit integer values.

USER INTERFACE

There are several inherent sources of delay in the recognizer, resulting in a pause between when a word is said and when the recognized word is displayed. The primary source of delay is due to the Language Model. Since words in the future can effect the probability of an extending word sequence, the Hypothesis Search does not generally make a firm decision on a given word until at least the following two words are tentatively recognized. This usually introduces a delay of about two seconds. On demonstration versions of the Tangora, the current best estimate of the partially decoded sentence is displayed in a separate window. The last few words displayed, which are subject to change, are called *infirm* words and are displayed at a low intensity. As decoding proceeds, and the Hypothesis Search determines that all possible alternative words are unlikely, it will make a final choice for the word, called the *firm* word. Displaying the words in this manner reduces the apparent delay to the user, since the estimate is usually the correct one and only the intensity of the word changes when it becomes firm.

This delay could be considered as an undesirable side-effect of the Language Model. However, the Language Model is essential for large vocabularies, and significantly improves recognition accuracy over the acoustics alone. For example, it would be quite difficult to correctly decode the following sentences without the Language Model.

"Twenty *two* people are *too* many *to* speak *to*."

"We need these *four* items *for four* weeks."

The Tangora has no difficulty with either of these sentences. It selects the proper *two*, *too*, or *to*, for example, based on the observed frequencies of the different spellings of the acoustically identical words, given the context.

Even with an initial vocabulary of 20,000 words, the Tangora cannot possibly know all of the names, technical terms, or acronyms that a speaker may want to use in dictation. One way to enter these words into a document is through *Spellmode*. In this mode, the recognizer is restricted to a small vocabulary consisting of letters, digits, and punctuation. Another way, most valuable for frequently used words, is by augmenting the vocabulary. We have developed an *Add Word* feature to allow personalization of the system's vocabulary for each user. New words can be added to the active vocabulary by providing the spelling through Spellmode or typing, and then saying the word to provide an example of its pronunciation.

The recognizer also supports the concept of *commands*. Some commands are acted upon by the recognizer, while others are passed to the speech application for execution. For example, saying *Spellmode* will cause the recognizer to enter that mode, and *EndSpellmode* will return the system to full vocabulary mode. In-line formatting requests, such as *NewParagraph*, *NewLine*, and *Erase*, are carried out by the speech workstation.

A *speech editor* has been developed for use in a remote speech workstation containing an Acoustic Processor connected via a telephone line to a Tangora recognizer. An existing what-you-see-is-what-you-get editor was modified to handle the special communications needed if speech is to be fully integrated into an application. For example, when the cursor is moved to a new text position, the Language Model requires the *left context*, or the previous two words in the document, to correctly compute the trigram probability for the next spoken word. The editor provides this information to the recognizer, which would have difficulty obtaining it without the cooperation of the user's application. It also supports the spoken formatting commands and additional speech functions. For example, it can query the Tangora to determine if a particular word is in the active vocabulary, or request that the Tangora add a new word to the vocabulary. The editor also displays the current best estimate of the recognized word sequence in the body of the document, with the infirm words displayed in a different font. Coupled with a terminal emulator program, it can be used to edit host VM/CMS files, and to create or respond to electronic mail through the CMS NOTE or similar facilities. Documents can be printed locally on the PC's printer, or remotely through a print file server on a PC network or on the host computer.

Speech can be added to existing applications without modification by the use of a *Speech Kernel*, which has been developed as a resident extension to DOS. This Kernel provides support for the remote Acoustic Processor functions and can be used by any application to interface to the recognizer. The Kernel places recognized words directly into the DOS keyboard buffer. This looks to the unmodified PC application, such as Personal Editor, as if the text had been typed. However, without the cooperation of the application, it is more difficult to provide all of the speech functions. For example, Language Model context changes due to cursor movement are currently detected by the Speech Kernel by reading the active screen display buffer, rather than being provided directly from the application. It is not always possible to determine the left context in this way, particularly if the cursor is in the upper left corner of the screen. Application specific templates are also supported by the Speech Kernel. This is most useful for macro expansion of recognized command words. For example, the *Erase*, *Save*, or *NewParagraph* commands can be mapped into different keystroke sequences for different editors or word processors.

TANGORA PERFORMANCE

The Tangora-20 can handle the 20,000 word vocabulary with essentially the same real-time response as the Tangora-5. The task-domain coverage of the 20,000 word vocabulary is a respectable 97.6%, much better than the 92.5% coverage of the 5,000 word vocabulary.

A comparison of the error rates of the Tangora-5 and Tangora-20 was made for *read speech* from seven different speakers, all members of the Speech Recognition Group. Six male speakers and one female speaker (M1-M6,F1) were used for these experiments. All except M6 were speakers of American English. The read speech consisted of sentences selected at random from documents written by a number of IBM employees. The sentences were on general business topics, and no documents by these authors were present in the databases used to derive the Language Model statistics. There were two sets of sentences selected, one containing only words in the 5,000 word vocabulary, and the other containing only words in the 20,000 word vocabulary. (The 5,000 word vocabulary is a subset of the 20,000 word vocabulary). The first set (S5) contained 50 sentences with a total of 884 words, while the second set (S20) contained 100 sentences with a total of 1,698 words.

Typical sentences from each of S5 and S20 with perplexities close to their average are:
"Managers should prepare themselves to ask the right questions about the individual's qualifications and to answer questions about IBM policies and practices."
"They installed the system over a single weekend, without unscheduled down time, and it has been working admirably since then."

RESULTS

Figure 1 compares the results of two experiments: decoding S5 with the Tangora-5, and decoding S20 with the Tangora-20. The average error rates are 2.9% and 5.4% respectively.

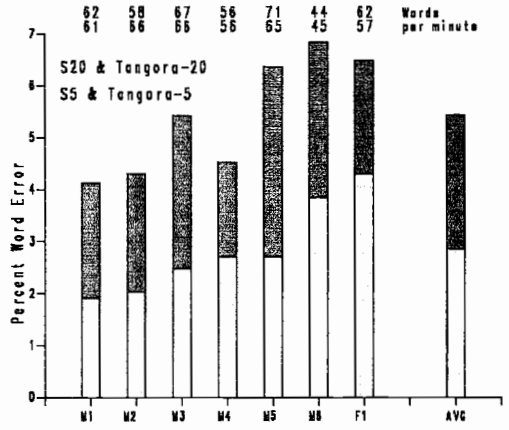

Figure 1. Errors for S5 on Tangora-5 & S20 on Tangora-20

The second task was the harder, of course, since not only does it have more acoustically confusable words, but also it has a larger perplexity. The perplexity [8] of a task is a measure of its difficulty based on information theoretic principles. For artificially constrained tasks it can be thought of as the average number of alternative words at each point. The perplexity of these two tasks as measured by their Language Models was 160 and 250 respectively.

In another experiment S5 was decoded with the Tangora-20, and the results are compared with the Tangora-5 error rate in Figure 2. The error rate increased on the average by 0.7% over that for the Tangora-5 since there were more acoustically confusable words in the larger vocabulary. Also the perplexity of the task as measured with the 20,000 word Language Model is slightly larger at 165.

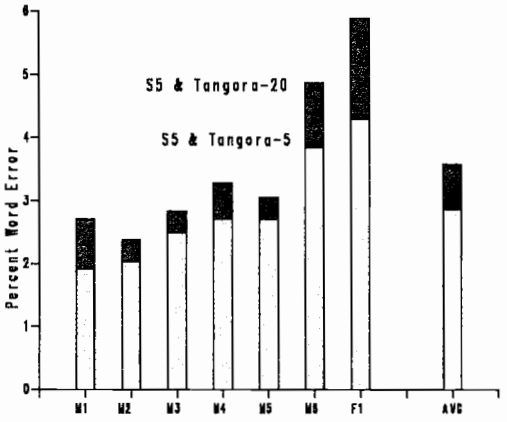

Figure 2. Errors for S5 on Tangora-5 & Tangora-20

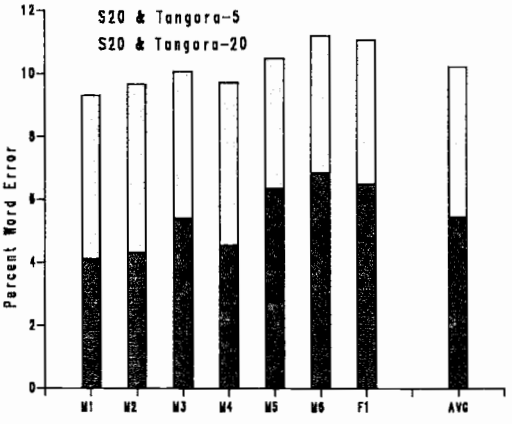

Figure 3. Errors for S20 on Tangora-20 & Tangora-5

Of the 1,698 words in S20, 1592 are also in the 5,000 word vocabulary. Hence the best possible error rate achievable when decoding S20 with the Tangora-5 would be 6.2%. Using the S5 average accuracy figure of 97.1% one might expect an accuracy rate of 91% (97.1*1592/1698) on the S20 task. The average

error rate actually increased to 10.2%, most likely due to the fact that the error produced by an out-of-vocabulary word can cause errors in adjacent within-vocabulary words by "derailing" the Language Model. Figure 3 compares the error rates when decoding S20 with the Tangora-20 and the Tangora-5.

The Tangoras are real-time for most speakers in the sense that it takes, on average, less elapsed time to recognize what has been said than it does to say the words. Figure 4 compares the decoding speeds of the Tangoras in seconds per second of speech for S5 on Tangora-5 and S20 on Tangora-20. The average decoding times are 0.84 and 0.94 seconds per second of speech, or 0.85 and 0.94 seconds per word spoken.

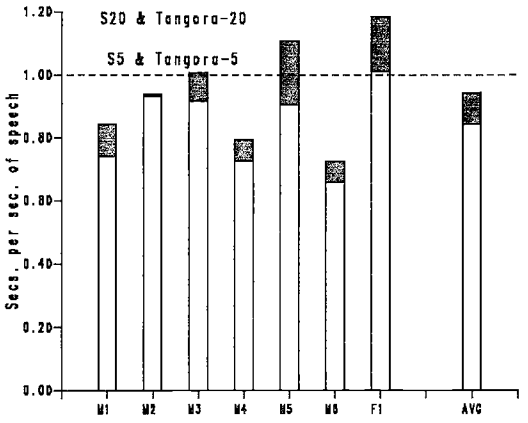

Figure 4. Comparison of Decoding Speed

CONCLUSIONS AND FUTURE WORK

The extension of the Tangora vocabulary to 20,000 words has significantly improved the useability of speech recognition for document preparation. It is expected that the addition of personal words to the base of 20,000 will provide almost complete coverage. As the physical packaging described here indicates, such capabilities should soon be economically feasible. The availability of spoken commands, and the ability to develop application programs that access the recognizer through the Speech Kernel, should greatly enhance the usability and customizability of the Tangora in different application areas.

Development of the Tangora is continuing in many areas. For example the PI subsystems can be reduced in size so that the Tangora-20 fits inside a PC-AT. The variations in performance among speakers as well as over time are being studied. Techniques are being developed to reduce the sensitivity to noise. There are also many questions relating to human factors that are being investigated.

REFERENCES

[1] A. Averbuch, L. Bahl, R. Bakis, P. Brown, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jelinek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman and P. Spinelli, "An IBM-PC based large-vocabulary isolated-utterance speech recognizer," *Proc. 1986 IEEE Int'l Conf. on Acoust., Speech and Signal Proc.*, pp. 53-56, Tokyo, Japan, April 1986.

[2] F. Jelinek, L. R. Bahl, and R. L. Mercer, "The design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Inform. Theory,* Vol. IT-21, pp. 250-256, May 1975.

[3] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE,* Vol. 64, No. 4, pp. 532-556. April 1976.

[4] L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis and R. Mercer, "Speech recognition of a natural text read as isolated words," *Proc. 1981 IEEE Int'l Conf. on Acoust., Speech and Signal Proc.*, pp. 1168-1171, Atlanta, Georgia, March-April 1981.

[5] A. Averbuch, L. Bahl, R. Bakis, P. Brown, J. Cohen, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, D. Fraleigh, M. Garrett, F. Jelinek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny and G. Shichman, "A real-time isolated-word speech recognition system for dictation transcription," *Proc. 1985 IEEE Int'l Conf. on Acoust., Speech and Signal Proc.*, pp. 858-861, Tampa, Florida, March 1985.

[6] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE,* Vol. 73, No. 11, pp. 1616-1624, Nov. 1985.

[7] G. Shichman, "Personal Instrument (PI) - A PC-based signal processing system," *IBM J. Res. Develop.,* Vol. 29, No. 2, pp. 158-169, March 1985.

[8] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.,* Vol. PAMI-5, No. 2, pp. 179-190. March 1983.

[9] J. Cohen, "Application of an adaptive auditory model to speech recognition," *J. Acoust. Soc. Amer.,* Supplement 1, Vol. 78, p. S50 (A), 1985.

[10] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer.", *IEEE Trans. Acoust., Speech, and Signal Proc.*, Vol. ASSP-34, no.3, March 1987.