Kurzweil VOICE^{TM}: Speech Recognition For Hands-Free Computing

Hong Z. Tan and David Horowitz
Center for Rehabilitation Technology
Department of Rehabilitation Medicine
Tufts University School of Medicine, Boston, MA

Aaron Kleiner and Joe Murphy
Kurzweil Applied Intelligence, Inc., Waltham, MA

ABSTRACT

Kurzweil VOICE^{TM} is a voice-activated personal computing system that gives you the ability to run word processing, spreadsheet, database, and popular MS-DOS software completely by voice. Some of its important features are: 40,000 pre-defined "speaker independent" English words that can be easily expanded to 50,000 by the user, high-level of voice recognition accuracy right out of the box, multi-tasking (i.e., the ability to run more than one application at a time), automatic speech adaptation, and the ability to run virtually any DOS-based applications completely via voice. Our speech recognition vocational services program provides engineering support and training for people with disabilities to use the Kurzweil VOICE to accomplish their jobs effectively. In an earlier paper (see Tan et al., CSUN'92 Proceedings) we discussed how we applied knowledge engineering to the Kurzweil Voice Report system to enhance a person's productivity on the job. In this paper, we discuss how we customized Kurzweil VOICE to support totally hands-free computing.

Introduction

The inaccessibility of computer systems to many people with disabilities is a major barrier to educational and employment opportunity. For those who are hands-impaired, and therefore unable to use computer keyboards, speech recognition technology has held the promise of overcoming this barrier. In order to be truly effective in this application, speech recognition systems should allow for hands-free use, have a very large vocabulary to permit dictation of general English text, should be relatively immune to extraneous noise and should be speaker independent. Speaker independence is the ability to with high accuracy recognize a user's speech without the user training the system on his or her voice. The Kurzweil VOICE system, developed by Kurzweil Applied Intelligence Inc. in Waltham, Massachusetts, is the first system to meet these criteria.

Speech Recognition Technology

Automated speech recognition occurs in three major stages: In the signal processing stage, analog sound signals are converted to digital signals. The stream of digital information is analyzed for acoustic-phonetic content and transformed into discrete segments of information that can be processed in the next stage. In the phonetic analysis stage, the information from the signal processing stage is compared with a database of known vocabulary items of sublexical units (word fragments). The output of this stage is a list of candidate words or other units which is examined more closely in the following stage. In the final recognition stage, words or other units which are recognition candidates are evaluated to determine which one is most likely to be the word spoken.

In the simplest systems, the selection of the best candidate is done strictly on the basis of phonetic analysis. In more sophisticated systems, other factors, such as syntax, may be used to eliminate some candidates and confirm others.

Speech recognition systems relying on a "pattern matching" approach to comparing input signals (spoken words) with stored reference patterns (vocabulary items) have serious limitation due to the large amount of variability in how individuals pronounce words without changing their meaning. Words may be spoken in many different ways; for example, slowly or rapidly, in different dialects, crisply or hoarsely, or slurred.

A variety of approaches have been explored to enhance the accuracy of computer speech recognition. One is the statistical approach in which probabilities of different pronunciations are estimated by examining many examples of previously recognized speech. This approach requires in-depth knowledge of the statistical relationships of all the words in the vocabulary and thus is computationally intensive, and the addition of new words to the vocabulary may be difficult. Another approach is based on recognition of phonemes, the basic sound combinations from which words are built. There are far fewer phonemes than words, and the difficulty of recognizing words is substantially reduced if phonemes can be recognized very accurately. The disadvantage of the phoneme recognition approach used-alone is that there is a high variability in the pronunciation of phonemes. In addition, phonemes are frequently slurred or omitted entirely by some speakers in certain circumstances. Finally, another approach is based on linguistics, and applies grammatical and syntactic rules in determining the most likely word used in a specific context. By taking into account the structure of language, the system can increase the accuracy of its decision making process.

Kurzweil's approach from the outset has been to apply a broad variety of methods, including techniques derived from statistical, phonetic and linguistic approaches as well as advanced pattern recognition. The Company uses and integrates proprietary versions of these approaches. Kurzweil's recognition technology is implemented as a collection of software "experts". These experts focus on different aspects of the recognition process, including acoustics, statistics, phonetics and linguistics. Results from all of the experts are combined to provide a final recognition. The weight given to different experts varies from zero to 100% according to the situation, and is determined by another software module called the Expert Manager, which coordinates the activities of the other software experts. If one expert is unable to recognize a spoken word in a particular situation (e.g., a noisy emergency room environment), the Company's multi-expert approach makes it more likely that another expert will be able to recognize the word accurately.

Recognition Accuracy

In order to evaluate the recognition accuracy of Kurzweil VOICE, we conducted independent tests with Kurzweil VOICE version 1.0.

Number of people tested: Six people participated in our recognition accuracy study. They were balanced in gender and with their experience with any speech recognition systems.

Procedure: Each user was first asked to practice with Kurzweil VOICE system for 15 to 30 minutes to learn the basic commands of Kurzweil VOICE and more importantly, to get comfortable with pronouncing discrete utterances in a consistent way. Since Kurzweil VOICE is a speaker independent system, a user can speak to it right out of the box without training it with his/her voice.

The user was then asked to record one sample per word from a list of 400 training words provided by Kurzweil VOICE. We required users to go through this voluntary training procedure so that our results can be directly compared to those obtained from other speaker-dependent speech recognition systems after adaptation. Depending on individual's speed, this adaptation procedure took 15 to 20 minutes.

Test Material: Each user was asked to read a passage of 287 utterances from the Introduction to the book "What it's like to be me". The following is the complete text used in our study.

> 'What it's like to be me was prepared during the International Year of Disabled People as a contribution by disabled children themselves. The basic idea was that this was to be their book, entirely their own words, entirely their own drawings, saying what disabled children themselves really felt. I hope that the book will communicate strongly to other children, to parents, to teachers, and to all of us.
>
> When I started the book, I suppose I expected it to be in some way miserable or depressing. But I found it the most genuinely happy book I have ever edited I had several disabled friends in childhood and so I never had some of the usual prejudices about disabled people. Then I trained as a speech and deaf therapist and it never occurred to me that I had any prejudices left. But I presumably had this one lingering prejudice that disabled people were in some way unhappier than the rest of us.
>
> Editing the book has, in fact, been quite a profound experience for me. Sometimes you can go through years of work without feeling that what you have done has changed you or given you a different dimension. But in this case the book has changed my outlook on life. Suddenly all my problems seemed to shrink and I knew somewhere deep inside that whatever life threw at me from now on I'd have the strength, like the children had the strength, to face the future.

The user was required to dictate the above text with proper punctuation and formatting commands. For instance, the very first six utterances made by each user were: introduction, cap-word, new-line, new-line, open-quote, and what.

Scoring Scheme: Every time Kurzweil VOICE made a recognition mistake, it was noted if the word spoken was in a 4-word alternatives list. (After the user makes an utterance, Kurzweil VOICE displays the Last Word Heard as well as 4 alternatives that it thinks are close to the Last Word Heard.) If the correct word appeared in the alternatives list, the user could use one of the take-one to take-four commands to correct it. We called these mistakes "secondary errors". Otherwise, the mistakes were called "hard errors". Note that the test material contained a made-up word "unhappier" that always resulted in a hard error.

Accuracy scores were computed in two ways. The lower bound of recognition accuracy was obtained by counting all correctly recognized utterances and dividing it by 287, the total number of utterances. The upper bound was obtained by adding secondary errors to all correctly recognized utterances and then dividing the sum by 287.

Results: All the recognition accuracy tests were done within 2 hours of each user's exposure to Kurzweil VOICE. Table 1 summarizes the results. Overall, the lower and upper bounds of recognition accuracy of Kurzweil VOICE after adaptation averaged 87% and 95%, which are much higher than those reported with any other large vocabulary speech recognition systems.

Table 1. Recognition accuracy results with two hours of exposure to Kurzweil VOICE.

Users SB and RR are individuals with severe disabilities who depend totally on a speech recognition system for their work as computer programmer/analyst. They achieved the highest recognition accuracies among all users tested. We feel that this is mainly because that these two individuals pronounced words more consistently than users who were less experienced with speech recognition systems. The other user, HZT, a native speaker of Chinese, who is also highly experienced with Kurzweil VOICE, didn't do as well because of her accent. Note that HZT can easily achieve a recognition accuracy of 90 to 95% with a well adapted voice profile. The results shown here were obtained from an out-of-the-box voice profile trained in less than two hours.

Hands-Free Applications

Kurzweil VOICE has many features that make it particularly suitable for people with disabilities. The two most important ones are hands-free and multi-tasking

By hands-free, we mean that you can do everything with your voice without ever touching the computer keyboard Once Kurzweil VOICE is loaded, it will keep running as long as the computer that runs it is left on. You can use a headset and say stop-listening and listen-to-me to turn Kurzweil VOICE "off" and "on", respectively. You can not only run various applications with your voice, but directly affect the way Kurzweil VOICE works. For example, you can say spell record to record voice samples of a word. This command lets you spell the word to be recorded using military alphabets (e.g., alpha for "a". You can also say capital to capitalize the next letter), record two samples of that word, and return to whatever you were doing. Another example is that you can say create-trigger to add a voice macro. This command lets you spell, with military alphabets, a trigger phrase (i.e., an utterance you would like to use) and its trigger text (i.e., the keystrokes to be sent to the appropriate application when you say the trigger phrase. You can include keystrokes like "control" or "F10" in your trigger text by simply saying control or F10).

By multi-tasking, we mean that you can run several applications at the same time with Kurzweil VOICE and Desqview. You can use three sets of commands to start, switch to, and shut down any applications. For example, you say start-WordPerfect to start WordPerfect, switch-to-WordPerfect to switch to WordPerfect from another application, and shut-down WordPerfect to release the link between Kurzweil VOICE and WordPerfect. Since Kurzweil VOICE maintains separate buffers, you can resume working in an application right where you left it after switching back to it.

The knowledge base that is supplied with Kurzweil VOICE includes predefined application voice commands for most of the WordPerfect commands. You can also interact with both DOS and Desqview entirely by voice.

To interface any other DOS-based applications, you will need to perform a number of interfacing steps that are described in details in Kurzweil VOICE User's Manual. In this paper, we'll list a number of commercially available software packages that we've successfully interfaced with Kurzweil VOICE. Note that except for CINTEX2, we've only implemented a partial set of application commands in the Kurzweil VOICE knowledge base for demonstration purposes. We can, however, always spell out any commands by saying the appropriate keystroke names.

CINTEX2, is a complete environmental control system that allows you to control up to 256 appliances, transmit over 100 trainable infrared patterns, and dial and answer your telephone. It is manufactured by NanoPac, Inc. in Tulsa, Oklahoma. We've interfaced Kurzweil VOICE with CINTEX2 so that it can be completely controlled with your voice. For instance, you can say start-CINTEX2 or switch-to-CINTEX2 to start or switch to CINTEX2 from another application. With CINTEX2 and Kurzweil VOICE, a person with severe disabilities can have complete control over his or her work and home environments.

To control home appliances with your voice, you need the "X-10 home control interface" and one 'X-10 appliance module" for each appliance you need to control (all of which are available from NanoPac). A pair of voice macros is created for each appliance. For example, you say turn-the radio-on and turn-the-radio-off to control your radio, you say turn-the-TV-on and turn the-TV-off to control your TV, etc.

To control infrared controllable devices with your voice, you need the "universal remote control" and the corresponding microprocessor from NanoPac. Once the universal remote control is trained with all the infrared control signals you need, you can use commands like TV-volume up, TV-volume-down, VCR-play, VCR-record, etc.

To control a telephone with your voice, you need a Hayes-compatible modem, a telephone amplifier, a universal module (all of which are available from NanoPac), and the "X-10 home control interface" mentioned earlier. You can pick up or hang up your telephone by saying answer-phone or disconnect-phone. You can dial a number by saying dial-phone, followed by digits, and then begin-dialing. CINTEX2 also allows you to set up a phone directory so you can speed dial a phone number. For frequently accessed phone numbers, you can create voice macros to combine all the necessary actions. For example, you can access your local time and temperature recording by saying time and temp.

Lotus 1-2-3 is a popular spreadsheet program that allows you to keep track of your expenses, create charts, and manage large amount of information (i.e., database). We've successfully interfaced Lotus 1-2-3 release 2.3 with Kurzweil VOICE.

AutoCAD is widely used for computer aided drafting. We've successfully interfaced AutoCAD release 10 with Kurzweil VOICE. Note that because AutoCAD uses the maximum amount of memory available, you should shut down other applications (except DOS) before you load AutoCAD with your voice.

Norton Utilities by Symantec is used for data recovery and protection, disk repair, speed and performance enhancement, security, etc. It contains many useful tools like save format, file find, system information, etc. We've successfully accessed many functions in Norton Utilities version 6.01 via Kurzweil VOICE. For instance, you can say change-directory to evoke NCD (Norton Change Directory). Note that some utilities, like Speed Disk, should not be run while any active application might access the disk. Since you need Desqview to run Kurzweil VOICE, you may not be able to use these utilities via voice.

ProComm is a communication software package made by Datastorm Technologies, Inc. in Columbia, MO. We've interfaced ProComm Version 2.4.2 with Kurzweil VOICE for (a) accessing a remote computing facility, and (b) exchanging files between the PC that runs Kurzweil VOICE and a remote computer, with a Hayes 2400 internal modem connected to a telephone line.

After saying start-ProComm, you can say alt followed by delta (for the alphabet "d") to bring up the Dialing Directory in ProComm. For the remote computing facility you frequently use, you can first create a Desqview macro (with appropriate time delays in-between sets of keystrokes) for bringing up the Dialing Directory, picking the appropriate entry in the Directory, dialing the number, and entering username and password. Then, a voice macro (such as connect-to mainframe) can be created to initiate the Desqview macro you've just created. If you frequently access several remote computers, you can create one voice macro for each remote computing facility.

Concluding Remarks

Kurzweil VOICE has a high-level of voice recognition accuracy, can be used in a totally hands free manner, and can run several applications at the same time. Kurzweil VOICE can meet your ever-changing educational and vocational needs with its ability to interface with virtually any DOS based applications. This makes it a highly viable and attractive alternative to keyboard input for computer access.

Acknowledgment

The first two authors are supported by a grant from the Department of Education Rehabilitation Services Administration (Grant #H128A01011-92). The first author wishes to thank Silvio Cianfrone and many people at NanoPac, Inc. for their technical assistance with interfacing Kurzweil VOICE and CINTEX2.

References

Hong Z. Tan, Robert R. Ebert, David M. Horowitz, and Corine A. Bickley (1992). "A Comprehensive Speech Recognition Service Delivery Model Illustrated with Case Studies". In Proceedings of CSUN'92, Technology and Persons with Disabilities, March, 1992.