ACHIEVING ACCEPTABLE ACCURACY IN A LOW-COST, ASSISTIVE NOTE-TAKING, SPEECH TRANSCRIPTION SYSTEM

Thomas Way, Richard Kheir* and Louis Bevilacqua
Applied Computing Technology Laboratory
Department of Computing Sciences
Villanova University
800 Lancaster Avenue
Villanova, PA 19085
{thomas.way, richard.kheir, louis.bevilacqua}@villanova.edu

ABSTRACT
Recent advancements in speech recognition technology and the widespread availability of inexpensive computers mean that real-time transcription in the classroom is becoming practical. This paper reports on the accuracy and readability achieved using a low-cost speech transcription system that assists deaf and hard-of-hearing students with note-taking. The design is presented for a system that attempts to balance issues of availability and cost with usability and accuracy to provide a viable and fully automatic alternative to a human signer or note taker. While this study focuses on empirical measures of design choices and accuracy, anecdotal results are also presented, with particular attention paid to the practical trade-offs between cost, accuracy and readability of a real-time transcription system.

KEY WORDS
Real-time speech recognition, assistive note-taking, accurate-enough transcription, low-cost assistive technology.

1. Introduction

Significant advances in computing power and speech recognition software that can achieve high accuracy have reached the mainstream [1], making speech transcription for use in the classroom as an assistive technology practical. Consumers are discovering that with moderate training and a modest investment, speech recognition systems such as Dragon NaturallySpeaking, IBM ViaVoice, MacSpeech iListen and recognition engines in the Windows XP and Vista operating systems can enable voice control of a personal computer, including good dictation accuracy under controlled conditions [2].

Not unexpectedly, there generally is a correlation between software cost and the accuracy, ease of use and customization for specific domains such as legal and medical dictation, with product prices ranging from $85 to $900 [2]. Cost can frequently be among the most decisive issues when a technology is to be used to assist persons with disabilities. As a result of these cost-related considerations, constraining the cost of a speech transcription system, while assuring that it provides sufficient accessibility and an accurate and readable transcript, is essential.

This paper reports on the design of a low-cost speech transcription system and the results of readability experiments. Experiments were performed to measure the accuracy as a way to evaluate the usability of this speech transcription system, based on results of earlier research. Design considerations are discussed that provide guidelines for implementing a low-cost, fully automatic system that is easy to use both for students and teachers.

2. Speech Transcription in the Classroom

While ubiquitous and affordable personal computing has made it possible to deploy automatic speech recognition (ASR) in the classroom, challenges remain [3]. Among these challenges are detecting and recognizing multiple speakers [4], incorporating visual cues [5], balancing the use of real-time automated speech text against the potential for distraction [6], providing sufficient accuracy in recognizing domain-specific jargon [4], configuring, training and deploying the ASR system for classroom use [7], and achieving acceptable accuracy through microphone selection, improved software and additional training of the ASR system [8].

Under ideal conditions, an ASR system can transcribe continuous speech more quickly than a human note taker, and with acceptable accuracy and training, making such systems a reasonable alternative for assisting deaf and hard of hearing students in the classroom [4]. Active research in ASR for college classrooms is being done by the Liberated Learning Project (LLP), among others [4,6,7,8,9]. The LLP has the goal of enabling students with various disabilities, including hearing impairment, to

* Richard Kheir is currently a software engineer with Hughes Network Systems in Germantown, Maryland.

maximize the benefits of the college lecture experience [10]. The LLP ViaScribe software performs real-time captioning, including ASR, of natural and extemporaneous speech. ViaScribe uses pause detection and phonetic spelling, among other techniques, to improve readability, and provides a less-accurate speaker-independent mode to accommodate multiple speakers [3]. The Rochester Institute of Technology produces the C-Print system produces a real-time transcript, although it requires a "captionist" who either re-voices to the ASR engine or physically keys in the transcription [9].

In the classroom, issues of sound reflectivity, air conditioning hiss, computer fan noise and other non-optimal sound conditions mean that the accuracy of a reasonably well-trained ASR system can easily fall into the range of 75-85% [8]. Accuracy rates of over 90% are possible for particularly consistent and clear lecturers using headset microphones [8], a rate that a significant majority of students find acceptable and useful [6]. The LLP originally devised a centralized ASR system to produce real-time captioning on a projection screen with post-lecture access to a transcription, although recent projects indicate that providing a more individualized experience where the transcript is delivered directly to each student may be preferable [8].

The use of ASR in the classroom requires that a speech profile be recorded, or trained, by the instructor, that a microphone be obtained for use both in this training and in class, and that a mechanism be provided for delivering the real-time speech transcription to students. The cost of an ASR system includes speech recognition software, a microphone and the computers used for transcription and delivery. Since accuracy is crucial, and instructor time is often in short supply, ASR software that can be trained in minimal time while achieving sufficient accuracy is needed.

3. Design of an Accurate, Low-cost System

Preliminary design, implementation and evaluation of a low-cost, real-time speech transcription were conducted at the Applied Computing Technology Laboratory at Villanova University [11]. Initially, the system focused on improving the accuracy of a widely-available and effectively free ASR system, the speech recognition engine accompanying the Windows XP operating system, by automatically extending the system dictionary using a custom software utility. A follow-up study determined that a usable transcript could be produced with approximately 30 minutes of ASR training performed by the instructor, and that delivering the transcript directly to student computers in a computer laboratory setting provided usable note-taking assistance [12].

The resulting Villanova University Speech Transcription (VUST) system was designed to minimize hardware and software cost, minimize the time an instructor spends training a speech profile, and maximize the readability of the resulting transcription. For purposes of this research, "usability" includes a quantitative measure of accuracy of the transcript and qualitative characteristics such as how well the transcript captures the intent and meaning of what was spoken and how useful it is to students who rely on it to assist their note-taking.

In a college classroom lecture setting, students receive information from multiple sources, including spoken word, projected slides, written material on a whiteboard and instructor-provided handouts. In designing the VUST system, the decision was made to provide the speech transcript directly to each student computer, either a laptop or desktop computer in a computer laboratory setting, rather than via a projected transcript as with the original LLP system [10]. In this way, it was hoped that the transcript could be referred to in much the same way as a handout, allowing a student to periodically refer to it rather than having to devote continuous attention so as not to miss anything. The VUST system is designed to be fully automatic, which improves upon systems such as C-Print which require a human assistant.

The VUST system consists of three major components: the speech recognition software, a dictionary enhancement tool, and a transcription distribution application. Figure 1 illustrates the VUST architecture, showing these major components and other elements of the system.

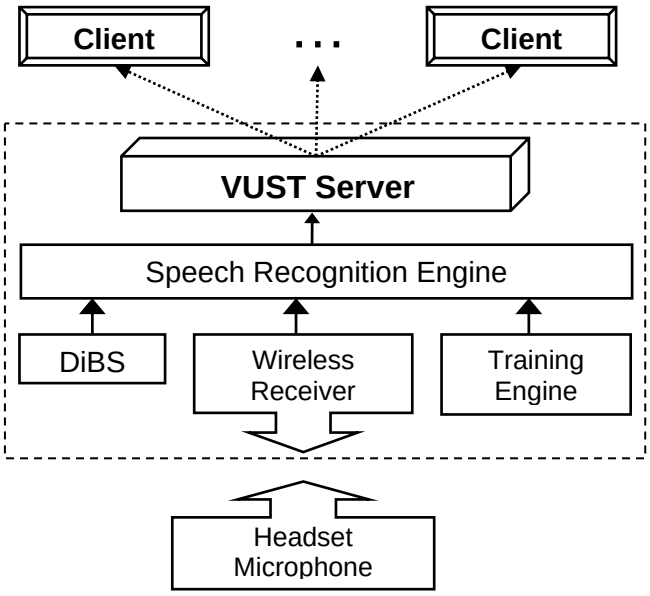

Figure 1. VUST speech transcription system design.

The dotted line in Figure 1 indicates the physical computer on which the speech recognition engine, VUST server application, wireless microphone receiver and other elements are located. One or more client applications can connect to the server, and a wireless headset microphone transmits speech to the server for

processing. The speech recognition engine uses the training engine and dictionary modification tool (DiBS), which is described below, to improve recognition accuracy.

3.1 Design guidelines and decisions

The general guidelines that were developed and followed for the VUST system are shown in Figure 2. The design of the speech recognition system required an ASR system that was affordable, accurate-enough and easy to set up and use. A variety of commercial and open-source products were evaluated and found to be similar enough in performance and ease-of-use that cost became the determining factor. Based on the goal of minimizing cost while achieving acceptable accuracy, the Microsoft Speech Recognition Engine (MSRE) was selected. The MSRE was selected due to the wide availability in academic institutions of the Microsoft XP platform, which includes the MSRE, effectively providing the ASR engine for our system at no additional cost. With the prevalence of student and instructor laptop computers, and computer laboratory classrooms, there was no additional cost for computers on which to run the VUST system.

Speech recognition engine
- Affordability – "free" or low cost
- Ease of use – easy to setup, train and integrate
- Accuracy – provides good-enough accuracy with minimal training (approx. 30 mins.)
- Supported – widely used and vendor supported
- Integration – recognition engine must provide programmer API or other easy integration into distribution application

Microphone
- Affordability – good cost for performance
- Interference – handles electronic interference
- Movement – allows instructor freedom of movement, wireless is best
- Noise – handles noise, unidirectional headset is best
- Recommended – by trusted, knowledgeable users
- Training – use the same microphone for training as will be used in the field

Transcription
- Distribution – direct to student laptop or desktop computer, not via projector
- Usability – transcript should be savable and scrollable
- Real-time – must be provide in close to real-time
- Ease of use – should be easy to acquire, such as Java applet or instant messaging client

Figure 2. Design guidelines for speech transcription system design.

Providing high quality speech input is necessary to achieve the best recognition accuracy, and the use of a wireless headset is a standard recommendation. Headset microphones provide consistent and close proximity to the speaker's mouth, and wireless transmission provides mobility for the speaker. A survey of available options online, including review of knowledgeable user feedback, resulted in the selection of the Nady Systems UHF-3 wireless unidirectional headset microphone. Particular weight was given to feedback from professional musicians and speech dictation users, who demand good performance from their microphones. Cost was again an overriding consideration, with appropriate wireless headsets available in the $90 to $400 range. The UHF-3 was selected as a cost-effective solution ($120-$140), with unrestricted movement, high directionality and good tolerance of electronic interference, such as that experienced in a college computer laboratory, being among the most important criteria when selecting a microphone for ASR [11].

The MSRE is trained by an instructor via a control panel included with the engine. The instructor reads from a selection of available text scripts into a microphone, enabling the recognition engine to learn to recognize the specific words as spoken by the specific instructor. The maximum level of training that was tested in our evaluation required less than one hour, with system setup, 30 minutes of script-based training, and 5 minutes to run the dictionary tool being all that was done.

3.2 Customization of dictionary

The MSRE relies on a static system dictionary for its basic recognition, with syntax rules built into the recognizer that phonetically match utterances with corresponding words. Secondarily, the recognizer uses words in a custom user dictionary in a similar way. Misrecognition is much more likely for words that are not contained in one of the dictionaries. Windows provides a user interface available to add individual words to the dictionary, and even by recording user-specific pronunciations of misrecognized words, but the process can be tedious and time-consuming. [11]

To improve recognition, particularly where a lot of topic-specific terminology and jargon is used such as in a college lecture, the Dictionary Building Software (DiBS) tool was implemented. The DiBS tool analyzes textual input, scanning for domain-specific terminology to add to the speech recognition system custom dictionary (i.e., "custom.dic"). DiBS parses an input file into words, filtering words below a minimum length threshold, that appear in a standard Unix system dictionary (approx. 10,000 words), and that already appear in the custom dictionary. The minimum length threshold of 6 characters limits the words considered to those with a higher likelihood of being domain-specific, which tend to be longer in length.

The key innovation of the DiBS tool is the ability for the user to easily add domain-specific terminology to the

MSRE custom dictionary in one, simple step. An instructor selects one or more documents, such as research papers or lecture notes, and the DiBS tool extracts jargon and customizes the dictionary, resulting in improved accuracy. The result is that the speech transcription will be of higher quality, and coupled with its low cost and ease of use, the overall system therefore will be much likelier to be used.

3.3 Distribution of transcription

The VUST consists of a text distribution server application and corresponding client application, both implemented in Java. The server and client are based on common chat server architecture, modified to accept input from the speech recognition engine and with client chat-back disabled. The design of VUST was kept minimal and straightforward to support a design goal of ease of use. Capture and acquisition of a lecture transcription had to be easy so that any instructor could deploy and use the system, and any student would find it easy to read and save the result. Java was selected as the implementation language to ensure portability across platforms, including Macs, PCs and Linux machines.

The VUST server receives the textual output of the recognition engine, and immediately forwards it to any client applications that are connected. The client application is a Java applet (Figure 3), embedded on a simple web page provided by the instructor, and automatically connects to the VUST server when the page is accessed. If the client fails to connect to the server, a message appears on the client indicating this failure.

In the sample of captured text in Figure 3, when brief pauses are detected, a period is inserted in the text, while longer pauses lead to the insertion of a paragraph break. In the last block of recognized text, even though the last sentence obviously contains some errors, it still maintains the intended meaning of the spoken sentence. This is typical of an acceptable form of recognition error.

In addition to presenting the live transcription of the lecture, the client also allows the student to export the transcription to a text file, copy and paste it to another program, or clear the current transcription from the screen. A pop-up dialog prevents the student from accidentally clearing a transcription in progress without first confirming the desire to do so.

Setting up and running the system involves ensuring that the instructor's computer is appropriately networked, connecting the wireless microphone receiver and putting on the wireless headset, activating the MSRE via the Windows Speech control panel, and starting the server application. Once the system is running, students can connect via a simple web page containing the client application. The instructor controls the location and content of this web page.

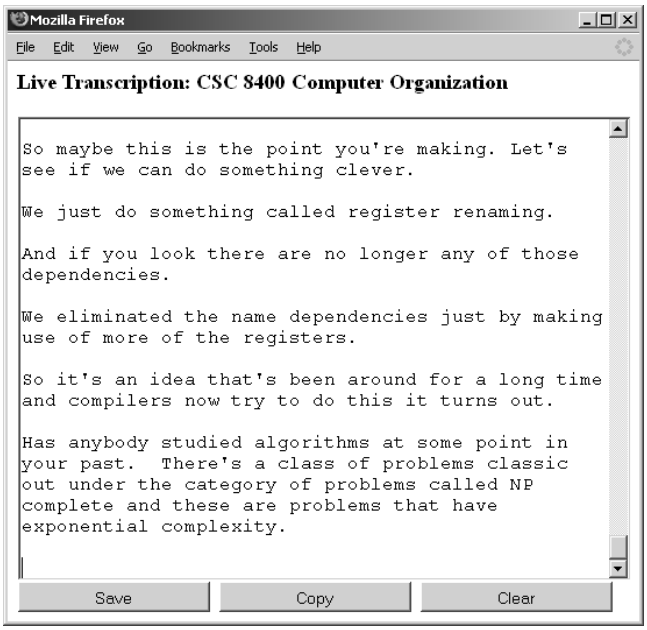

Figure 3. VUST Transcription Client applet.

4. Experimental Evaluation and Results

Experiments were conducted to measure the quantitative accuracy achievable under varying degrees of training and the qualitative acceptability of the resulting transcript as a way to provide note-taking assistance in a classroom setting.

4.1 Methodology

The VUST system was tested using prepared lecture notes in a real classroom setting. The VUST server was run on the instructor's laptop computer with the wireless microphone receiver connected to it. The instructor wore a wireless headset microphone, which is the same microphone that was used to train the system. Moderate training of the recognition engine (30 minutes) was performed, and the DiBS tool was used to customize the dictionary with related content.

Transcripts were captured using VUST and also recorded in digital audio form to WAV files which were later transcribed by hand for comparison. The lecture material was terminology-rich, taken from undergraduate and graduate level courses in computer science, delivered by a computer science professor. No restrictions were placed on instructor movement, and student interaction was encouraged. When possible, the instructor repeated student questions or comments to ensure the interaction was included in the transcript. Measures of accuracy were performed using a text file comparison tool called "DiffDoc" (softinterface.com) and then manually analyzed for verification.

4.2 Achieving acceptable accuracy

Research into the readability of speech recognition transcription has determined that an accuracy of at least 90% is required to make a transcript usable [5], and some believe that accuracy of 98% or better may be needed to maintain the meaning and intent of the content [14]. Previous experiments revealed that VUST could provide accuracy that was sufficient to students at 88%, and good at 90% and above [11].

A 90 minute lecture, consisting of 9,783 words, was transcribed using the VUST system as described above. The transcription output was saved to a text file and also transcribed manually from the captured digital audio file for comparison. The instructor then analyzed the transcript and identified all misrecognitions, within reasonable constraints (e.g., singular vs. plural and homonym misses were allowed when the meaning was intact, while obviously incorrect recognition or anything that hurt the meaning was marked as incorrect). The automatic and manual transcriptions were then compared for accuracy. Sections of the transcript were classified based on their speech content, as: roll-call (list of names or otherwise discontinuous speech), planning (assignments, dates, and general classroom business), discussion (interaction including student discussion), and lecture (continuous instructor speech). This procedure was repeated two additional times by inputting the recorded digital audio file into the VUST system, and the results of all three experiments were averaged. It is worth noting, however, that recognition was very stable and that there was very little variation and the three test runs.

Usability of the resulting transcription was measured by reading the transcript and in effect grading it as if it were a student report summarizing the content of the lecture. Using the system as described, the transcripts that were produced were deemed to be passable as class notes with only minor editing, such as inserting paragraph breaks. The transcript was not verbatim, and certainly not perfect, but was sufficient to convey the content and meaning of what has been presented to support the goal of assisting with taking notes.

Not surprisingly, the best recognition accuracy was achieved with prepared lecture, resulting from the MSRE preference for continuous speech. Table 1 summarizes the quantitative results obtained using the VUST. Overall accuracy was 89%. Planning, lecture and discussion were all generally consistent with this average, with roll-call scoring well below (61%). The ranges of accuracy, measured on a per paragraph basis, for each classification of speech, are provided. These ranges support the characteristic of most speech recognition engines, such as the MSRE, that continuous speech is preferred and produced better results. Distribution of the paragraph accuracy measures was roughly Gaussian for each classification of speech, centered on the average accuracy.

Table 1. Comparison of VUST recognition accuracy of four classifications of speech content.

Classification	Words Correct	Total Words	Overall Accuracy	Accuracy Range/¶
Planning	635	758	84%	79-88%
Lecture	6329	6925	91%	85-96%
Roll-call	155	254	61%	58-63%
Discussion	1592	1846	86%	82-90%
TOTAL	**8269**	**9783**	**89%**	**58-96%**

The low recognition accuracy (61%) of roll-call speech was not unexpected. A student name can be a form of domain-specific terminology all to itself, and are not likely to be found in the static system dictionary. Planning speech scored next lowest (84%), due to its disjoint, bullet-item nature, which at times lacks the continuous flow that the MSRE prefers. Discussion and lecture speech were both recognized at better rates, deemed usable by the instructor and students who participated. Discussion was at the borderline of being usable (86%), while lecture material was above at least one minimum estimate of usability [5] at 91% accurate.

4.3 Qualitative feedback

Qualitative feedback was elicited from students to evaluate the usability of the transcript and to identify strengths and weaknesses of the current system. Students participated in the experiments by completing a survey following the lecture. Although only one of the 30 students in the test group was deaf, the hearing students were asked to be as objective as possible in evaluating the usability of the VUST system and transcript. Students felt that having the transcript on the computer screen in front of them made it easy to refer to, convenient to use, and not distracting. Typical comments were that the transcript was “pretty good” and that it “got the meaning,” although some felt that “it made some weird errors” and “the formatting was hard to read.” Overall, students viewed the transcript produced as “helpful” and believed that it could provide valuable assistance in taking notes.

These qualitative results support other work done in this area that found that a transcript is only usable when the original meaning is maintained and the formatting of the transcript is not a distraction [5]. For the deaf student participant, the experience of real-time transcription that was usable was exciting and very helpful. The deaf student had not relied on assistive technology or sign interpreters in the past, and for the first time found himself fully engaged in a lecture, raising his hand to contribute to classroom discussion in real-time as a result of the VUST transcript in front of him.

5. Conclusion

The VUST system shows significant promise as an affordable and beneficial assistive system to make the classroom more inclusive for deaf and hard of hearing students. Although this research focused primarily on the technical design issues, the quantitative and qualitative assessments indicate further development and experimentation is warranted. The system demonstrates that acceptable speech transcription accuracy of 85%, or even 90% or greater with the addition of dictionary customization, can be achieved by a low-cost and easy to use system. Distributing the transcript in real-time can greatly benefit deaf and hard-of-hearing students, the instructors who teach them, and potentially even hearing students, although it requires additional custom programming beyond the capabilities of commercial ASR software. The use of a wireless headset microphone is important enough that a modest expenditure will be money well spent.

The use of the speech recognition engine in the Windows operating system minimizes the cost of the system, although popular commercial alternatives such as Dragon NaturallySpeaking could fit within a reasonable budget. The VUST system was designed to be distributed free of charge, with the microphone and access to a Windows-based laptop the only outside factors. Migration to the Windows Vista platform is an alternative, with claims that the MSRE in Vista is improved over its XP counterpart. According to vendor supplied information, the speech recognition engine in Vista has improved accuracy, an easier to use interface, continuous adaptive training so that recognition accuracy should improve with use, and continued support for software developers who want to integrate ASR into their applications. [13]

Technical and cost issues that must be addressed when considering commercial software solutions include whether or not there are licensing fees associated with distributing the software to students, whether the software provides API or means for programmers to connect to it, control the engine and distribute the transcription, and what long-term support will be available for the product. Personnel costs cannot be overlooked, and unlike some other available ASR transcription systems, VUST does not require a human captionist to be "in the loop."

Future development of VUST will include improvements to pause detection and the insertion of punctuation and spacing, phonetic transcription for words that have low recognition confidence to benefit readability of misrecognized words, and design of a centralized repository system for domain specific terminologies and speech profiles to enable easier and more effective use of the system. The evaluation of recent advances in recognition technology, including the MSRE in Windows Vista, and other cost-effective commercial and open-source options is planned.

References

[1] D. Pogue, Telling your computer what to do. *The New York Times*, March 1, 2007.

[2] Consumer Search, Voice recognition software consumer report. May 2007, available online at http://www.consumersearch.com [accessed Jan. 3, 2008].

[3] K. Bain, S. Basson, A. Faisman and D. Kanevsky, Accessibility, transcription, and access everywhere. *IBM Systems Journal, 44(3)*, 2005, 589-603.

[4] C. Davis, Automatic speech recognition and access: 20 years, 20 months, or tomorrow? *Hearing Loss, 22(4)*, 2001, 11-14.

[5] R. Stuckless, Recognition means more than just getting the words right: Beyond accuracy to readability. *Speech Technology*, Oct./Nov. 1999, 30-35.

[6] A. Hede, Student reaction to speech recognition technology in lectures. In S. McNamara and E. Stacey (Eds.), Untangling the Web: Establishing Learning Links, *Proc. of the Australian Society for Educational Technology (ASET) Conference*, Melbourne, July 2002.

[7] K. Bain, S. Basson and M. Wald, Speech recognition in university classrooms. *Proc. of the Fifth International ACM SIGCAPH Conference on Assistive Technologies*, ACM Press, 2002, 192-196.

[8] M. Wald and K. Bain, Enhancing the Usability of Real-Time Speech Recognition Captioning Through Personalised Displays and Real-Time Multiple Speaker Editing and Annotation. *Universal Access in Human-Computer Interaction. Applications and Services*, Lecture Notes in Computer Science, Springer, 2007, 446-452.

[9] P.M. Francis and M. Stinson, The C-Print Speech-to-Text System for Communication Access and Learning. *Proceedings of CSUN Conference Technology and Persons with Disabilities*, California State University Northridge, 2003.

[10] Liberated Learning Project, Coordinated by Saint Mary's University, Halifax, Canada. Available online at http://liberatedlearning.com [accessed Dec. 30, 2007].

[11] R. Kheir and T. Way, Improving speech recognition to assist real-time classroom note taking. *Proc. of Rehabilitation Engineering Society of North America (RESNA 2006) Conference*, Atlanta, USA, June 2006, electronic proceedings, 1-4.

[12] R. Kheir and T. Way., Improving Access to Computer Science Education with Speech Recognition. *Proc. of 12th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2007)*, Dundee, Scotland, June 25-27, 2007, 261-265.

[13] R. Brown, Talking Windows: Exploring new speech recognition and synthesis APIs in Windows Vista. *MSDN Magazine, 21(1)*, 2006.

[14] Type Well, Speech recognition in the classroom. Online article, October 2006, available online at http://typewell.com/speechrecog.html [accessed Jan. 4, 2008].