Improving Speech Recognition to Assist Real-time Classroom Note Taking

Richard Kheir, MS, Thomas Way, PhD
Applied Computing Technology Laboratory
Department of Computing Sciences, Villanova University

ABSTRACT

Speech recognition systems have advanced to the point where they are a viable option for providing note taking assistance for deaf and hard of hearing students. College lectures, which frequently contain domain-specific or uncommon terminology, provide a challenge for these systems that typically rely on a dictionary of common words to guide recognition. This paper reports on a prototype text analysis software tool, and some general configuration techniques, that can improve the ability of an affordable and off-the-shelf speech recognition system to assist deaf and hard of hearing students with note taking in a college classroom setting. Some specific technology choices are discussed and the results of a preliminary evaluation of the text analysis tool are presented.

KEYWORDS

Speech recognition, ASR, assistive technology, note taking, college classroom.

BACKGROUND

Advances in affordable portable computing technology have led to wider availability, making it possible to deploy automatic speech recognition (ASR) in the classroom, although challenges remain (1). The ability of ASR systems to transcribe continuous speech faster than a note taker can write, with reasonable accuracy and minimal training, make them a viable option to assist deaf and hard of hearing students with note taking (2). Obstacles to relying on ASR for note taking include recognizing multiple or random speakers (2), synchronizing and incorporating visual cues (3), balancing real-time automated speech text (AST) against the potential for distraction (4), insufficient accuracy in recognizing domain-specific jargon (2), configuring, training and deploying the ASR system for classroom use (5), and achieving acceptable accuracy through microphone selection, improved software and additional training of the ASR system (6).

Active research in ASR for college classrooms is being done by the Liberated Learning Project (LLP), among others (2,4-6). The LLP has the goal of enabling students with various disabilities, including hearing impairment, to maximize the benefits of the college lecture experience (7). Significantly, the LLP has partnered with IBM to develop the ViaScribe software that is specifically designed for real-time captioning, including ASR, of natural, extemporaneous speech. ViaScribe improves readability by detecting pauses in speech and inserting sentence and paragraph breaks, provides phonetic spellings when the recognizer is uncertain, and even has a less-accurate speaker-independent mode to accommodate multiple speakers (1). Accuracy of reasonably well-trained ASR systems typically is better than 75-85% in classroom lecture settings, with rates over 90% for particularly consistent and clear lecturers (2,6), a rate that a significant majority of students find acceptable and useful (4). The common scenario of a centralized ASR system producing real-time captioning on a projection screen with post-lecture access to a transcription has been used successfully in the classroom (6), although a more individualized approach often may be preferable (1,4,6).

This paper presents the results of a pilot study conducted at the Applied Computing Technology Laboratory at Villanova University (actlab.csc.villanova.edu) to evaluate the impact of our Dictionary Building Software (DiBS) utility and the accuracy of a portable, cost-effective, laptop-based ASR system designed to augment note taking by deaf and hard of hearing students in the college classroom.

RESEARCH OBJECTIVE

The goals of this study are to quantify the impact of a new dictionary customization software tool on ASR accuracy in a college lecture setting, and to develop guidelines for deploying it within a cost-effective, individualized ASR system for assisting deaf and hard of hearing college students.

METHODS

This study measures the effectiveness of the DiBS utility to improve the recognition accuracy of the Microsoft Speech Recognition Engine (MSRE). The engine was prepared and tested using five training scenarios: untrained, minimally trained, moderately trained, moderately trained with a customized dictionary, and moderately trained with a customized dictionary and selected customized pronunciations. Tests were performed using spoken lectures containing terminology-rich material from undergraduate and graduate courses in computer architecture, totaling approximately 3,700 words or 30 minutes of continuous speech. The lectures were conducted in a classroom by a computer science professor wearing a wireless headset microphone, using a very clear and consistent speaking style, and were digitally captured to WAV files. To enable valid comparison, these digitized lectures were then replayed to the MSRE running on a university-issued laptop, under five training scenarios, with the AST output captured into a Microsoft Word file. Objective measures of accuracy were made using a free text file comparison tool called DiffDoc (softinterface.com) by comparing the AST with a transcription of the original lecture. Results of the comparison tool were analyzed manually for verification.

The experiments were conducted using an ASR system designed to be affordable, accurate and easy to set up and use. 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. The Nady Systems UHF-3 wireless unidirectional headset microphone was selected as a cost-effective solution ($120-$140), with unrestricted movement and high directionality being key considerations when selecting a microphone for ASR (7). The maximum level of training that was tested required less than one hour, with 30 minutes of script-based training, 5 minutes to run the DiBS tool, and 10 minutes of additional training to record pronunciations of domain-specific words. The DiBS tool analyzed a number of text files containing the content of technical papers and lecture notes related to the subject matter of the test lectures. Custom pronunciations were recorded using the MSRE training interface for approximately 10 domain-specific words that the MSRE had difficulty recognizing. The DiBS tool culls terminology from input files by selecting only those words not found in a dictionary of common words, and then appending the selected words to the dictionary (i.e., “custom.dic”).

RESULTS

Table 1 shows the accuracy and usability of the results of recognition. Accuracy improved with additional training, with marked improvements when going from an untrained to a minimally trained system (from 75% to 88% accurate) and with the addition of a customized dictionary and pronunciations to a moderately trained system (from 91% to 94%). The recognition accuracy varied greatly (plus or minus 5-10%) depending on the prevalence of terminology that was not found in the default ASR dictionary. Adding terminology from the domain of the lecture helped, and additional recording of pronunciations of specific terminology that the recognizer still misrecognized helped more.

Insert Table 1 here: Comparison of recognition accuracy.

Usability of the resulting AST was measured by reading the transcript and in effect grading it as if it were a student report summarizing the content of the lecture. This more subjective usability of each transcript was judged broadly to be: poor, fair, sufficient, good, very good, excellent. Even with minimal training, the results were passable (sufficient), although they required careful reading and some editing to make them usable as notes. With moderate training, transcripts were usable (good) as class notes with only minor editing, such as inserting paragraph breaks. Although very good usability was achieved with the addition of some customized pronunciations, excellent usability was not achieved in any of the scenarios, reinforcing the need for continued research in speech recognition technology (1).

DISCUSSION

Customizing the dictionary of an ASR system with unfamiliar terminology is effective at improving accuracy. The DiBS tool provides an efficient means to automatically cull such domain-specific jargon from large amounts of text and customize the ASR. Effective use of an ASR system in college classrooms requires not only accuracy and usability, but a means to re-train individual systems. Using the Speech Recognition Profile Manager Tool (microsoft.com), a speech profile can be imported or exported, making possible distribution of the profile, along with custom dictionaries for specific topics, via a central repository such as a university or department web site. The DiBS tool enables faculty to create a customized dictionary that improves SR accuracy, reducing the time required for training a profile.

Future work includes plans for development of a prototype distributed application that automates and integrates training and customization of SR systems, development of other add-on improvements to available off-the-shelf systems that will improve their usability for deaf and hard of hearing persons in classroom and business environments, and facilitation of real-time, networked delivery of AST, as in the ViaScribe system.

REFERENCES

1. Bain, K., Basson, S., Faisman, A., and Kanevsky, D. (2005). Accessibility, transcription, and access everywhere. IBM Systems Journal, Vol. 44, No. 3, pp. 589-603.
2. Davis, C. (2001). Automatic speech recognition and access: 20 years, 20 months, or tomorrow? Hearing Loss, 22(4), pp. 11-14.
3. Stuckless, R. (1999). Recognition means more than just getting the words right: Beyond accuracy to readability. Speech Technology, Oct./Nov. 1999, pp. 30-35.
4. Hede, A. (2002). Student reaction to speech recognition technology in lectures. In S. McNamara and E. Stacey (Eds.), Untangling the Web: Establishing Learning Links. Procs. of the Australian Society for Educational Technology (ASET) Conference, Melbourne, July 2002.
5. Bain, K., Basson, S., and Wald, M. (2002). Speech recognition in university classrooms. Procs. of the Fifth International ACM SIGCAPH Conference on Assistive Technologies. ACM Press, pp. 192-196.
6. Wald, M. and Bain, K. (2005). Using automatic speech recognition to assist communication and learning. Procs. of the 11th International Conference on Human-Computer Interaction, Las Vegas.
7. Liberated Learning Project (2006). Coordinated by Saint Mary's University, Halifax, Canada, http://www.liberatedlearning.com [accessed Jan. 2, 2006].

Richard Kheir, Thomas Way
Applied Computing Technology Laboratory
Department of Computing Sciences
Villanova University, Villanova, PA 19085
(610) 519-5033, richard.kheir@villanova.edu, thomas.way@villanova.edu

GRAPHICS

Table 1: Comparison of recognition accuracy and usability.

Table 1. Comparison of recognition accuracy, range of accuracy, and usability.

Description	**Accuracy**	**Range**	**Usability**
Untrained	75%	64-83%	poor to fair
Minimal training (default script, 10 minutes total)	88%	78-93%	sufficient
Moderate training (3 additional scripts, 30 minutes total)	90%	81-96%	good
Moderate training, customized dictionary	91%	83-96%	good
Moderate training, customized dictionary, customized pronunciations	94%	86-98%	very good

Alternative text description for Table 1: Comparison of recognition accuracy.

Table depicts the percentage of accuracy for speech recognition for the system with it is untrained, minimally trained, and with moderate training, including the effect of a customized dictionary and customized pronunciation. Accuracy improves from 75% for the untrained system to 94% with moderate training, a customized dictionary and pronunciations. The table also shows the range of accuracy for each level of training, which is a measure of how accurate recognition was on a paragraph by paragraph basis, and the comparative usability of the resulting transcription.