

Speech Transcription Services

Dimitri Kanevsky, Sara Basson, Stanley Chen, Alexander Faisman, Alex Zlatsin

IBM, T. Watson Research Center, Yorktown Heights, NY 10598, USA

kanevsky@us.ibm.com

Sarah Conrod

Cape Breton University, Sydney, N S, Canada

sarah_conrod@capebretonu.ca

Allen McCormick

ADM solutions, Dominion, NS, Canada

adm@admsolutions.ca

Abstract

This paper outlines the background development of “intelligent” technologies such as speech recognition. Despite significant progress in the development of these technologies, they still fall short in many areas, and rapid advances in dictation have actually stalled. This paper proposes semi-automatic solutions for smart integration of human and intelligent efforts. One such technique involves improvement to the speech recognition editing interface, thereby reducing the perception of errors to the viewer. Some other techniques described in the paper include batch enrollment techniques which allow users of speech recognition systems to have their voices trained by a third-party, thus reducing user involvement in preliminary voice-enrollment processes. Content spotting, a tool that can be used for environments with repetitive speech input, will be described. Content spotting is well-suited for banks or government offices where similar information is provided on a daily basis, or for situations that have repeated content flow as in movies or museum tours.

1. Introduction

As bandwidth for web and cell-phone applications has improved, there has been a corresponding increase in audio and video data available over the web. This proliferation of data has created new challenges in accessibility including the need for better methods of transforming spoken audio into text. The creation of textual transcripts ensures that multimedia materials are accessible not only to deaf and hard of hearing users, but also for users who are “situationally disabled” with limited technical resources. For example, users with low bandwidth transmission can read text streams when full bandwidth video is not an option. Written audio transcripts are also prerequisites for a number of other important processes, such as translating, summarizing, and searching.

Human transcription of speech-to-text is expensive and requires trained experts, such as stenographers. Through the use of a human intervention called “shadowing,” “parroting,” or “re-speaking,” automatic speech recognition has been used for live transcription. In these methods, speakers train speech recognition software to recognize their voices. These speakers

then “shadow” the speech of untrained speakers, similar to the process used in simultaneous translation. Studies suggest that the accuracy levels of software trained on such skilled voices make “shadowing” a viable tool for live subtitling [8]. Effective, shadowing is a task that requires skill and training; therefore, it is not widely available or affordable.

Methods for automatic speech transcription are steadily improving, with error-rates dropping by as much as 25% per annum on specific data domains [9],[10],[11]. Nonetheless, full transcription of audio materials from all domains remains a distant goal. For example, current automatic transcription accuracy for broadcast news is approximately 80% [7],[12]. There is an obvious gap between current automatic performance and the professional requirements for captioning. As a result, expensive human intervention, rather than speech technology, is currently used in commercial captioning. This costly process poses another problem: most audio and video information remains untranscribed, untranslated, unsummarized, and unable to be searched. Speech captioning is becoming an increasingly important issue in our information society. This paper suggests ways to provide speech captioning by bridging what speech automation technology can currently handle, and what can best be handled through human mediation. It also represents a brief overview of some aspects of the paper [6]. The challenge before us is to utilize the human component most efficiently, and cost-effectively, while simultaneously improving speech automation technologies. This paper focuses on distributed post-editing as a viable mechanism to counter shortcomings in speech technology. The paper also describes several other methods to address current shortcomings of speech recognition, such as content spotting, batch enrollment and automatic speaker identification.

2. Using existing speech recognition technology for accessibility

2.1. The Liberated Learning Project

The Liberated Learning Project provides an example of how people have used automatic speech recognition to caption lectures. In “Liberated Learning” courses, instructors use a specialized speech recognition application called IBM ViaScribe. When used during class, ViaScribe automatically

transcription is typically outsourced into locations with relatively low labor costs.

iv) *Editing*

Editing of speech recognition errors is only efficient for transcription tasks if the number of errors is relatively low (below 20%) [1], otherwise, it is more efficient to transcribe audio manually. Off-line editing can be provided by specially trained people who trained in methods that speed up editing.

v) *Alignment of audio and text*

Webcasts with video and audio should be aligned with transcribed text in order for a viewer to be able to associate a transcription with the related video. If a transcription text is obtained via an automatic speech recognition system, then the audio is already aligned with a decoded text. Tools for editing of the decoded text that is aligned with the audio should be organized in such a way that corrected text remains aligned with the audio. If the transcribed text was produced manually, automatic means for text-audio alignment of the text should be used or the alignment of the audio with the text should be done while a writer transcribes the text (by mapping small

segments of the audio that is being played with the pieces of text that the writer produces).

Dollar signs in blocks in Figure 1 show relative “costs” of the corresponding processes. For example, human transcription with a stenographer is generally more “expensive” than shadowing. These situations were observed for some off-line CaptionMeNow transcription processes that were tested jointly by IBM and partners of LL in Nova Scotia, Canada. The diagram in Figure 1 does not cover all possible hypothetical transcription processes that would require a more elaborate analysis.

While speech recognition cannot be used alone to reliably caption random audio, our goal is to determine whether a hybrid of speech recognition plus limited human intervention can nonetheless reduce the time and costs associated with captioning. In our transcription work, we found that one hour of audio could be transcribed in 3-4 hours using a speech recognition system described in [13], plus editing. The same audio using manual transcription performed by experienced writers required 6 hours of effort.

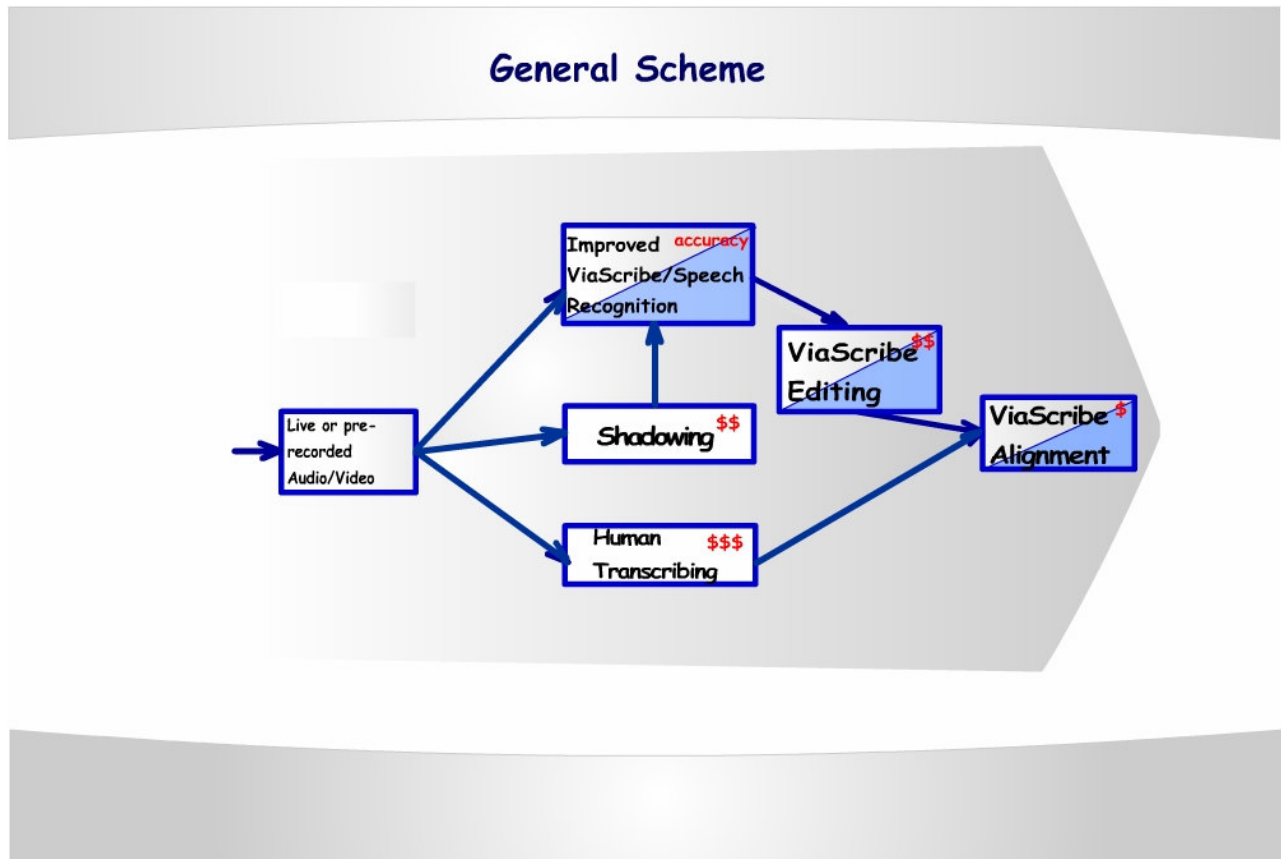


Figure 1. Paths for Captioning

2.3. Museum Applications

Another successful application of ViaScribe speech recognition technologies was recently demonstrated at the Alexander Graham Bell National Historic Site (AGBNHS) in Nova Scotia, Canada [4]. Conclusions emerging from this application are presented in the following section. This Baddeck Liberated Learning Showcase was launched in 2004

and was a project designed to explore how the use of real-time transcription could enhance interpretative talks within a National Historic Site. Due to their specific nature, museums provide an ideal environment to employ speech recognition technology for several reasons. One reason that speech recognition tools are particularly well suited to museum settings is because people that are hired to interpret museum stories tend to be highly efficient public speakers.

3. Editing innovations

Speech automation technologies can provide transcription, but editing work is typically necessary to ensure a high level of accuracy. In one case, as in interactive dictation, an author of the dictated speech corrects decoding errors. In another case where speech is uninterrupted, as in real-time captioning, editors correct speech recognition errors for other speakers. Editing will slow down the process, however, since multiple hours of editing might be required to perfect a one hour transcription. In some cases, real-time editing will be necessary to ensure that the correctly captioned material is immediately available. Multiple editors can be used to accelerate the editing process, but this presents another challenge: how to efficiently coordinate multiple editors, and ensure that the final product is provided in real time and appropriately synchronized. Off-line transcription of audio that is stored in webcasts or library archives requires a different approach for editing which depends on the accuracy requirements, which then depends on time and availability of speech recognition and human editing resources.

There are several issues that should be developed in order to increase the user interface editing efficiency.

- “Work Sharing” enhancements will enable multiple editors to be privy to what other editors have already done, and what sections other editors are currently working on.
- A special approach can be used to provide a “quasi-real” time transcription, i.e., transcription of speech that should be created with no more than 5-7 seconds delay. Such delays of transmission are acceptable in some TV broadcasts of live events. Video is delayed until the transcription is produced and then transcription is inserted into a video stream. In this “quasi-real” time scenario, an expert, for example, a person who does shadowing can also mark speech recognition errors by clicking on them. The marked words are then split between editors who work on audio and textual segments that are assigned to them. This assignment of textual segments to different editors can be done by marking these segments with a different color.
- Editing tools will also include multiple input mechanisms such as a mouse, keyboard, touch screen, voice, pedal movement, or a head-mounted tracking device that is placed on a person’s head that moves a cursor to a point where the user is looking on the display. Additionally, different aspects of the task might be better handled by different interface tools. For example, “spotting” errors could be easier to handle with a head-mounted device while repairing phrases might be best handled via a keyboard in the “quasi-real-time” scenario that was described above. Interactive dictation, for example, when the speaker edits errors himself, can be done mostly orally/visually, perhaps with just a pointer to highlight what is wrong and needs to be repeated. A quick mouse highlighter and oral feedback might be the fastest and most efficient model. The most efficient routing mechanism will be determined empirically based on task assessments.
- For off-line tasks such as transcription of audio archives, editing can be distributed randomly, based on the availability of editors. Alternatively, the editing work can be distributed hierarchically. For example, the “first pass” review can be provided by “phase 1” editors, with a

certain level of skill; the “second pass” review may be distributed to “phase 2” editors with a different level of skill. Some editors may have unique expertise in particular terminology therefore this phase of editing will be routed to these more specialized editors.

- The user interface of choice needs to be adaptable to the user, and change dynamically depending on what types of errors are displayed. The different error types result in the display of different kinds of tools to facilitate editing these errors. The tools will be multimodal, allowing the editor to choose the best modality for a given problem. A word that is identified as incorrect, for example can then lead to a display of an alternative word list, so the editor can just click on the word that he wants to correct; or the reader can repeat the misrecognized word orally, to try for a better recognition the second time. In other cases, it will allow the editor to use handwriting and to overwrite the incorrect word. The interface suggests user modalities that would fit the most suitable user for this type of error. The network of editors will also provide data that allows the system to draw conclusions about editor habits and preferences, thereby improving and customizing the editor interface across the network.
- In current speech recognition dictation systems, there is no “look ahead” or “look behind” mechanism. A particular error may occur repeatedly, and each instance of the error needs to be corrected individually. A more efficient system would automatically correct similar errors across the entire transcript. The correction of an error should automatically modify language probabilities and thereby allow automatic correction of other similar errors.

4. Usability enhancement: batch enrollment

Standard speech recognition enrollment requires a speaker to read prepared stories aloud. The recorded audio is then used to train the speaker model. A typical enrollment process involves reading one story and usually takes 20-30 minutes. Some users may read three or four different prepared stories to improve their acoustic models. Reading more than 3-4 stories rarely improves decoding accuracy. Further improvement in the user model may occur when the user corrects speech errors that occurred during dictation.

This type of enrollment process has the following deficiencies. Many people are not willing to spend hours training speech recognition systems to achieve higher accuracy. Some are not willing to read even one enrollment story. These types of examples can be observed in the medical arena. Medical doctors produce a lot of written material, they usually dictate into a tape that is then transcribed by medical transcription services. Such transcription services usually involve manual transcribers. A similar situation was observed in the Liberated Learning project where only a handful of lecturers, “early adopters” of the speech recognition technology, agreed to enroll their voices into the speech recognition systems.

It also becomes difficult to enroll the system when there is no access to the speaker, which is typical for much of the audio data that is archived and available. Another problem with the standard enrollment process is that it records speech when a speaker reads a story rather than recording free conversational dialogue. An acoustic model that is created from a story read by a user may not capture certain

5. Summary

This paper outlines the background development of “intelligent” technologies such as speech recognition. Despite significant progress in the development of these technologies, they still fall short in many areas, and rapid advances in areas such as dictation have actually stalled. Reduced efforts toward developing speech recognition technology for dictation have inadvertently affected the development of speech recognition for high bandwidth applications. This negatively affected the accessibility area, where high bandwidth, large vocabulary speech recognition technology was critical and required for applications like webcast and lecture captioning. Alternative solutions like stenographic transcription address some of these gaps, and innovative efforts such as the Liberated Learning Consortium advanced the state of the art for providing transcription in classrooms using speech recognition.

While the technology still needs to evolve, we have presented a very general concept for filling some of the gaps left by technology-driven solutions of speech recognition. We have proposed semi-automatic solutions – smart integration of human and intelligent efforts. One such technique involves improvement to the speech recognition editing interface, thereby reducing the perception of errors to the viewer. Other efforts underway make the technology easier to use for the speaker. Batch enrollment, for example, allows the user to reduce the amount of time required for enrollment.

Clever selection of applications can also bypass some of the technical shortcomings. Applications that have repeated content flow, such as movies or museum tours, can more easily be transcribed, and even translated. For these areas, content spotting technology, where key words or phrases trigger the right content, can bypass any inaccuracies in the speech recognition technologies.

Important issues remain unresolved and should be the topic for future research. For example, there is a need to develop adaptive user interfaces for error correction that dynamically present the best tools for rapid correction and also make errors obvious. This includes development of an optimized scheme for error correction. For example, identification of errors in one part of the text can automatically trigger identification or correction of similar errors elsewhere in the transcript. Another important area that should be developed is the infrastructure for distributing process components across a range of services to the appropriate technologies and human providers, and finding the best path for efficient, cost-effective allocation of work.

6. References

- [1] Bain, K., Basson, S., Faisman, A. and Kanevsky, D. 2005. “Accessibility, transcription, and access everywhere”, *IBM Systems Journal*, 44(3): 589-603.
- [2] Bederson, B., 1996. “Audio Augmented Reality: A Prototype Automated Tour Guide.” Bell Communications Research”. *Proceedings of CHI '95*: 210-211
- [3] CaptionMeNow, 2004. http://www-306.ibm.com/able/solution_offerings/captionmenow.html
- [4] Conrod, S. 2005. “Enhancing Accessibility in Interpretative Talks”, *Proceedings of the Conference: Speech Technologies: Captioning, Transcription and*

- Beyond, IBM, June. <http://www.nynj.avios.org/Proceedings.htm>
- [5] Kanevsky, D., Basson, S. and Fairweather, P. 2004. “Integration of Speech Recognition and Stenographic Services for Improved ASR Training”, U.S. Patent 6,832,189.
- [6] Kanevsky, D., Basson, S., Faisman, A., Rachevsky, L., Zlatsin, A., Conrod, S., 2006, “Speech Transformation Solutions”, to appear in *Special Issue on Distributed Cognition*.
- [7] Kim, D., Chan, H., Evermann, G., Gales, M., Mrva, D., Sim, K. and Woodland, P. 2005. “Development of the CU-HTK 2004 Broadcast News Transcription Systems,” *ICASSP05*, Philadelphia, PA, March.
- [8] Lambourne, A., Hewitt, J., Lyon, C., Warren, S., 2004. “Speech-Based Real-Time Subtitling Services”, *International Journal of Speech Technology*, 7: 269–279. <http://homepages.feis.herts.ac.uk/~comrcml/IJSTarticle.pdf>
- [9] Nahamoo, D. 2006. Personal communication.
- [10] Pallett, D. 2003. “A Look at NIST’s Benchmark ASR tests: Past, Present, and Future”. http://www.nist.gov/speech/history/pdf/NIST_benchmark_ASRTtests_2003.pdf
- [11] Saon, G., Povey, D. and Zweig, G. 2005. “Anatomy of an extremely fast LVCSR decoder”, *proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4-8, pp. 549-553.
- [12] Saraclar, M., Riley, M., Bocchieri, E. and Goffin, V. 2002. “Towards automatic closed captioning: Low latency real time broadcast news transcription”, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA.
- [13] Soltau, H., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Zweig, G., The IBM 2004 Conversational Telephony System for Rich Transcription, *ICASSP 2005*.