# Transforming Access to the Spoken Word

*Douglas W. Oard*

College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD, 20742 USA
oard@umd.edu

## Abstract

For thousands of years, the written word has held a special place in our lives. In part, this results from two key characteristics: durability and searchability. Over the past several decades, the spoken word has gradually acquired those characteristics. In our lifetimes, it seems reasonable to expect that trend to continue, and indeed to accelerate, as improvements in automatic speech recognition begin to enable large-scale access to spontaneous conversational speech. This paper identifies four fundamental challenges that must be overcome if we are to leverage this remarkable new capability for the greatest benefit, briefly describes one project that is exploring this new frontier, and then concludes by looking toward future research on this important problem.

## 1. Introduction

Humans have an innate ability to produce and comprehend speech; writing and reading are, by contrast, acquired skills. Among the most affluent societies, most of our citizens learn to read. Literacy is still far from universal in many areas of the globe, however. Perhaps even more importantly, the written record of our civilization has for the most part been produced by the relatively few of us that choose to tell our stories in that form. Humans are storytellers, optimized through evolution for sharing their stories verbally. Why, then, has the written word achieved such prominence? Fundamentally, there are two reasons. For tens of thousands of years, the spoken word was ephemeral; once spoken, stories lived only in the minds of those who heard them. The written word, by contrast, has a far greater degree of permanence, ranging from decades for common materials to millenia for the earliest written records that have survived to the modern era. Over the last century, that advantage has evaporated. Some recordings that I made two decades ago are still usable, and we have some basis to believe that with careful curation speech that is digitized today will still be usable millenia hence. In a digital world, speech and writing are equally durable.

The key to explaining the enduring prominence of the written word must therefore lie in the second advantage that writing has historically enjoyed; that which once was written can later be found when it is again needed. Libraries, and today digital libraries and information retrieval systems, evolved from simple arrangements of manuscripts on a shelf to complex systems that provide content-based access to enormous collections of hyperlinked Web pages. The simplest of these techniques apply equally well to speech, but approaches based on manual arrangement and indexing are simply not scalable. Automated indexing of speech is now possible, but when compared with processing electronic text, processing speech is slow, brittle, and inaccurate. Automatic speech recognition systems that run in real time seem fast when compared to systems that were available just a few years ago, but tokenization—the equivalent process for electronic text—is about six orders of magnitude faster. As a result, Google crawls and retokenizes three billion Web pages every few weeks. With the same processing power, we could acquire and re-recognize a comparable number of spoken words every few thousand years. Clearly, we have a ways to go before we are as facile at manipulating speech as we now are at manipulating electronic text. Of course, Moore's law will eventually get us part of the way there, and evolutionary improvements in processor architectures and recognition algorithms could reasonably be expected to close some of the rest of the gap.

It therefore seems reasonable to ask what will happen once speech can be converted to tokens almost as easily as we presently tokenize electronic text. The short answer seems to be that it will change everything. Present systems are brittle in the sense that they need to be trained using hand-transcribed examples that are representative of the ultimate application. But a handful of applications would suffice to enable access to enormous quantities of recorded speech. Already people have built workable systems for dictation, broadcast news, telephone calls, meetings, and oral history interviews. Today's largest operational systems index over 10,000 hours of speech.[1] The content that could potentially be indexed is vastly larger, however. The British Library oral history collection alone contains about 250,000 hours of recorded interviews, and SingingFish (a service that indexes Web audio based solely on metadata) has indexed more than 35 million audio streams.[2]

Yogi Berra, a famous American pundit, once said, "it's tough to make predictions, especially about the future." But if we are willing to together envision a future in which automatic speech recognition is sufficiently efficient to be affordable at enormous scales and sufficiently robust to be applied to a broad range of potentially content, then only one real problem remains. We don't really know what to do with it. Automatic speech recognition gives us the potential to transform access to the spoken word, but realizing that potential is another challenge entirely. In the next section, I describe four important facts about speech retrieval that take us beyond what we understand from our experience with searching the written word. Section 3 then describes the way in which we are exploring those issues in the MALACH project. Finally, Section 4 concludes the paper with some thoughts about directions for future research that would build on the work that we already have underway to advance the research agenda for transforming access to the spoken word.

[1] http://speechbot.research.compaq.com
[2] http://singingfish.com

## 2. Spoken Word Collections

The first steps in any new area naturally build on what came before; so it has been for speech retrieval. Broadcast news was the first large-scale source to become tractable for automatic speech recognition, and it proved to be a useful place to begin. Many of the standard text retrieval test collections contain written news articles; extending similar evaluation techniques to spoken news stories proved to be straightforward. Speech recognition accuracy initially posed some challenges, but those challenges were overcome through improved retrieval techniques (notably, document expansion) and continued improvements in recognition accuracy. Today, this is generally seen as a solved problem; broadcast news recognition with acceptable error rates (generally, well below the 40% word error rate at which the best retrieval techniques begin to degrade) are available for a number of languages, and a number of research prototypes (e.g., Informedia) and commercial systems (e.g., Virage) have been built.

Broadcast news is often recorded under studio conditions by professional announcers with excellent diction; most recorded (and recordable) speech poses much greater challenges for automatic speech recognition systems. Recently, however, it has become possible to automatically transcribe spontaneous conversational speech with an accuracy comparable to what has already been shown to work well for broadcast news retrieval. This opens up a vast array of possible applications:

**Broadcast programming.** In addition to broadcast news, this category includes interviews, talk radio, sports, entertainment and advertising.

**Scripted stories.** This genre is characterized by planned speech for which automatic alignment to an existing script may be possible. Examples include books on tape, poetry readings, and theatrical productions.

**Spontaneous storytelling.** This genre is characterized by spontaneous speech; examples include oral history interviews and recordings of folklore.

**Incidental recording.** This is by far the most diverse and challenging category, spontaneous speech that was produced principally for some purpose other than creating a recording. Examples include classroom lectures, political speeches, courtroom oral arguments, business meetings, and telephone calls.

Most of these potential applications have a recognizable constituency of real users with real information needs, and we could easily add several more to this list. In the coming years we can reasonably hope to be able to build systems that can automatically transcribe the speech that is present in each case with reasonable accuracy. Our success as a community with broadcast news retrieval might lead us to conclude that speech retrieval and text retrieval are not all that different; once you have run speech recognition, what you are left with is a text retrieval problem. Natural language exhibits considerable redundancy, so "bag of terms" approaches to retrieval are inherently robust in the presence of moderate word error rates. This tempting proposition overlooks four key challenges that are masked by the familiar nature of news stories:

- We don't usually tell stories in ways that are easily divided; it makes sense to search for broadcast news stories as separate entities, but such unambiguous divisions into retrievable units are not a natural characteristics of most human speech.
- Much of what we say would be of little or no value to anyone in the future. Put another way, the information density in most of the speech that could be recorded is quite low. That's not true in broadcast news, where professional announcers (usually) seek to make effective use of limited air time. We might see similar problems in text-based forms of conversational media (e.g., chat rooms or personal email), but these applications have not yet been well studied. As a result, searching conversational speech is a needle-in-a-haystack problem at a scale that is unprecedented in mainstream research on text retrieval or in research to date on speech retrieval.
- It is not really clear what we would do with a good ranked list if we could make one. In broadcast news retrieval, a few extracted terms can make for a good "headline" for a story, and a list of headlines can be browsed quite quickly. What should we do for conversational speech? Perhaps we can pinpoint some "hot spots" within a recording where the user could begin to listen. But listening to even a few brief segments would risk interrupting the rapid iterative convergence on a good query that is a hallmark of effective interactive searching in many applications.
- How will our society cope with the new capabilities that we provide. Since the dawn of history, we have treated the spoken word as ephemeral. For example, here in the USA there are laws against recording telephone conversations without the knowledge of the participants in most situations; such laws exist solely to perpetuate a social construct that technology long ago rendered obsolete. Searching recorded news raises few concerns among our citizens, but we are about to move into a realm where these issues will rise to be on a par with the technical issues that we are better positioned as a research community to address.

If we are to make progress on these challenges, we need access to large collections of recorded speech, automatic speech recognition systems that can transcribe that speech with acceptable accuracy, and real users that have real information needs in those collections. The next section describes how that is being done in one project.

## 3. The MALACH Project

The goal of the MALACH (Multilingual Access to Large spoken ArCHives) project is to advance the state of the art for access to large multilingual collections of spontaneous conversational speech by leveraging an unmatched collection assembled by the Survivors of the Shoah Visual History Foundation (VHF). The VHF collection contains 116,000 hours of interviews that were conducted in 32 languages with nearly 52,000 survivors of the Holocaust. By the end of next year, the entire collection will have been digitized (to 180 TB of MPEG-1 video) and manually indexed using an extensive controlled vocabulary and within-interview name authority control. To this remarkable collection of "found data," our colleagues at the IBM TJ Watson Research Center, the Johns Hopkins University Center for Language and Speech Processing, and Charles University and the University of West Bohemia in the Czech Republic have added automatically created transcripts for nearly 1,000 hours of English and Czech using speech recognition systems that achieve word error rates below 40%.

A 10,000-hour subset of the VHF collection was hand-segmented into topically-coherent segments with an average length of 3 minutes (about 400 words), and the 640 hours of English speech recognition transcripts were drawn from that subset. We used those transcripts as the basis for a small test collection to explore segment-oriented retrieval. Seventy search topics have been created from actual requests that had been submitted in writing by potential users of the VHF collection, and relevance judgments for the approximately 10,000 segments in the test collection were created at the University of Maryland for 28 of those topics using a search-guided assessment methodology. Fourteen topics were independently assessed; the 44% agreement on positive judgments compares favorably with similar (55%) results for test collections created for the Text Retrieval Conferences using a pooled assessment methodology.

Recent initial experiments with this test collection at the University of Maryland, the Johns Hopkins University Applied Physics Laboratory, and the IBM TJ Watson Research Center have yielded some interesting results. Speech retrieval has indeed proven to be challenging, with the best present results (from IBM) yielding mean average precision values slightly below 0.10, corresponding to an average of 3.4 relevant documents in the top 20. Viewed another way, searchers would need to listen to about 15 minutes of audio on average before hearing the first relevant segment. These results are still quite preliminary, but they tend to reinforce our belief that the four issues identified in the previous section will indeed be serious challenges.

The first question to ask is whether the manual segmentation we have used reflects a real use scenario. In order to answer this question, we worked with VHF to study a group of 8 high school teachers that searched some of the hand-segmented interviews for clips that they might show in their classes. Not surprisingly, these teachers sometimes preferred a different span for their segments than had originally been created by the indexer. This tends to reinforce our belief that the end user must be the ultimate judge of what the passage boundaries should be. We are therefore considering a shift to unsegmented search (initially using a sliding window approach), evaluated using passage overlap measures similar to those that were tried in the TREC-2003 High Accuracy Retrieval of Documents (HARD) track.

A second question is whether the relatively low retrieval effectiveness that we have observed reflects low information density or if is is merely indicative of the relative lack of sophistication in our initial experiments. Another possibility is that these results may reflect some as-yet uncharacterized weakness in our implementation of search-guided relevance assessment. We are presently manually transcribing every known relevant segment for three topics; results with those accurate transcripts will help to guide our future work on recognition and retrieval. We have found that searching manually written segment summaries (which are also available for the 10,000-hour subset) yields a mean average precision above 0.3, so if the right words are there, we know that we can find them fairly often. We will soon know whether such words were spoken, and that understanding will help to guide our future work with these materials.

The question of what to do with the resulting ranked list is still somewhat open, but the work of others (notably at BBN and MITRE) with broadcast news suggests that using speech recognition results for automatic text classification can provide a useful degree of description. We and our colleagues at the IBM TJ Watson Research Center have trained two types of text classification systems using 3,000 manually transcribed segments. The best present systems achieve a balanced F measure above 0.25, and we believe that we may be able to achieve substantial improvements by focusing on a smaller category set.

Finally, there is the question of how we might share the unique resources that we are creating. At present we have an information retrieval test collection and over 400 hours of manually transcribed spontaneous conversational English, Czech, Russian, and Slovak that can be used to build speech recognition systems. The challenge here is not technical; rather, our principal obligation is to provide appropriate protections for the integrity of the stories that have been told and some aspects of the privacy of those who have told them. This challenge can certainly be overcome, but doing so will require a dialog between our research community and those who curate these types of collections.

## 4. Conclusions

Of course, no single project will turn up everything that we need to learn. There has also been some work with voice mail (notably, at AT&T), recorded meetings (at Berkeley and elsewhere), lectures (at Cornell, MIT, and the Tokyo institute of Technology) and political speeches (at Fraunhofer IMK and as a joint effort of the University of Colorado and Michigan State). Work with the spoken word at the Text Retrieval Conferences (TREC) has evolved to include a broader range of content than broadcast news as part of the VideoTREC evaluation, and the HP Labs SpeechBot system is exploring scalability issues on content that extends beyond broadcast news as well.

What new frontiers remain? One intriguing direction would be to explore retrieval from ubiquitous personal audio. The potential for augmentation of human memory using such a system could be immense, but the social challenges (e.g., protection of privacy and intellectual property) may ultimately prove to be equally large. More measured steps might include focusing on a broader range of historical materials (e.g., courtroom oral arguments and presidential phone calls). Each new collection brings new users, and with them new user needs. This, in turn, can inspire new research directions. When all else is equal, we will probably learn the most from the applications that are the least like what we have seen before. So while we may not yet know quite where we are going, we at least have some idea how to recognize the best directions in which to head. On May 6 of this year, a group of us will gather at a workshop following the Human Language Technologies conference in Boston to consider these questions. I invite you to join us there as we seek to envision our common future.

## 5. Acknowledgments

The work reported here reflects the contributions of the entire MALACH team and insights gained through discussions with members of the US/EU Digital Library Working Group on Access to Spoken Word Collections,[3] and the organizing committee for the HLT/NAACL workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval. Comments by Bill Byrne, Martin Franz, and Bhuvana Ramabhadran on an earlier version of this paper are particularly appreciated. This work has been supported in part by NSF IIS Award 0122466, NSF CISE Research Infrastructure Award EIA0130422, and IBM through a Shared University Research award.

[3] http://www.dcs.shef.ac.uk/spandh/projects/swag/