

Transforming Access to the Spoken Word

Douglas W. Oard, University of Maryland

Why focus on the Spoken Word?

In this brief paper, I've sought to elucidate a portion of what I think we should consider advocating as a way of steering future investments in information access research. To be useful, information must be found, and it is on that prerequisite that I will focus here; our discussions at the workshop should, of course, also include the question of how information will be used, and how search and use can best be integrated.

People discover information in many ways, but what we typically call "search" is distinguished by three characteristics: it is intentional, it is a monolog rather than a dialogue, and it places the initiative with the receiver. Although we sometimes refer to systems that we build as "search engines," search is best viewed as a process pursued by human and machine together, rather than as a job to be done by a machine. Machines bring two unique capabilities: speed and reliability. Humans bring three: intention, intelligence, and decision. Without machines, the scale of human search would be severely constrained. Our challenge is therefore to envision, create and assess ways in which machines can support this human activity.

The spoken word is innately human; almost everyone learns to speak, and to understand other speakers. We are story tellers by nature, and for most of our time on this planet, the spoken word has been the means by which we shared our understanding and passed our cultural heritage and technical knowledge to succeeding generations. A few thousand years ago, the introduction of writing changed all that. The written word now occupies a privileged position, for three reasons: (1) the written word can be stored for extended periods and then reproduced unchanged, (2) we have developed techniques and technology to help us find words that have previously been written, and (3) we can teach people to produce and interpret written words. The first and third of these are now true of the spoken word as well; at 2 megabytes per hour, a single 250 GB hard drive can now hold more speech than I will produce in my lifetime.

Erasing the one remaining barrier; search, offers the potential to transform the way we live and work for at least three reasons. First, access to the spoken word is empowering; reading and writing can be taught, but we often do not do that as well as we would wish, thus permanently limiting the potential contributions of many of our fellow citizens. Second, effective access to the spoken word can dramatically accelerate innovation. We depend on information technology to compress the innovation cycle, seeking to move information nearly instantaneously from those who create it to those who need it. But as long as what moves must be written, nothing will go anywhere until someone writes it down. The vast majority of what we write is spoken long before it is written. Until we come up with a way of sharing our thoughts while they are still in our mind, improving access to the spoken word is probably the most important thing that we can do to further

accelerate the innovation cycle. Finally, the spoken word offers the potential for expression in ways that the written word simply cannot match.

Of course, speech will not displace writing; both offer unique advantages. But we now have the potential, for the first time in human history, to place the written and the spoken word on an equal footing. It is hard to imagine any other investment that could bring greater benefits to our society.

Setting the Research Agenda

A sustained investment over two decades has yielded some effective ways of searching spoken content. Useful accuracy (at about a 15% word error rate) has been demonstrated for radio and television news broadcasts and for personal dictation, and accuracy that is sufficiently good to support term-based searching (<40% word error rate) has been demonstrated for some specialized applications (e.g., voice mail, recorded telephone calls, and oral history interviews). Automated searching has been explored in the context of broadcast news materials, and fully integrated interactive search systems have been demonstrated for a few applications (e.g., radio Webcasts and news broadcasts). Development of speech recognition systems for new domains remains expensive, however; achieving a 40% word error rate for some specific combination of genre and language typically requires several person-years of effort, an investment of perhaps a million dollars. Continued investments in further improving accuracy are most certainly well justified, but investments in breadth and robustness are equally important. The integration of audio and video processing seems to offer substantial promise in this regard; humans do better under difficult acoustic conditions when they can see the speaker, and if we develop the right techniques it seems reasonable to expect that the same will be true of our machines.

Speed is another place where focused research investments could offer transformational potential. Processing spoken words in a way that will allow them to be searched in the future is far slower, about a million times slower, than processing written words. Modern Web search engines processes about one trillion words every week; the largest present index of spoken words on the Web has processed a bit under a billion words in two years. There is much that we can do with systems that run at present speeds; but there is much more that we could do if the processing were much much faster.

Far more words are spoken each day than are written, and one key consequence of this is that many of the gems we seek to find are likely to be buried in a far larger sea of distracting content than is the case with present systems that are oriented towards written materials. Addressing this challenge will require a broad view; spoken word collections pose challenges, but they also offer opportunities. We already know how to automatically create some types of metadata that can help to address this challenge (e.g., speaker identification and tracking), and initial work in a number of other areas seems promising (e.g., emphasis and emotion detection). Moreover, emerging techniques for behavior-based characterization of information content offer the potential to integrate

estimates of features such as source authority and affective reaction that can further improve the utility of automated search systems.

The diversity of human language tends to balkanize the information space, and the increasing penetration of information technology throughout the world serves to exacerbate this effect. This has important consequences for commerce, security, and the “digital divide,” so it is important that we mitigate the effect to the extent possible. Effective cross-language search technology has been demonstrated with spoken word collections, but we must augment the automated capabilities of those tools with effective broad-coverage speech translation if we are to realize the full potential of this technology. Present speech translation systems rely on a limited domain of discourse and the opportunity for correction dialogs to achieve satisfactory performance. Open domain speech translation may soon be within our reach, however, leveraging the redundancy that is naturally present in human expression and the potential to exploit multiple modalities (e.g., spoken, written and graphical) to convey a useful degree of understanding. We will need to make exploratory investments before we can clearly discern the full set of opportunities, but the potential impact of this research is enormous and some starting points are now becoming apparent.

Realizing this bold vision will require that we integrate three key research strategies. At the core of the effort will be a series of interdisciplinary research projects that bring together the full range of perspectives needed to explore one or more driving applications (e.g., archival access to recorded meetings, creation and use of personal lifelines, or improved access to specific types of cultural heritage collections). We must commit substantial resources to achieve effective collaboration at the scale needed for application-centered research, but this strategy is the key to shaping a research agenda that matches needs with opportunities.

While “big science” fills a crucial shaping role, individual initiative is the wellspring of innovation. We must, therefore, balance our investment between a few large projects that will illuminate the way ahead and a broad array of smaller research efforts to create the technology base on which everything else is built. Finally, we must recognize that as we create new capabilities, we must help our society learn how to employ the technology we create in beneficial ways. Many of our social conventions and a substantial part of our legal framework are based on a shared understanding that the spoken word is ephemeral. We are about to change that permanently, and the consequences of that change on public policy and on individual behavior will undoubtedly be pervasive. It is incumbent on us to help our fellow citizens understand the nature of the tools that we will give them, and to help explore the policy implications of these new capabilities. This, then, is the third focus of the integrated research strategy that we envision.

The research described above promises to benefit every corner of our global village, and achieving that potential will require investments from a broad range of stakeholders. No one nation can effectively attack the challenges of linguistic diversity alone. No one agency has a charter so broad that it could single-handedly address all of the implications of this technology for commerce, security, and society. And no one program could

possibly envision everything that needs to be done. Our approach must be iterative and multifaceted. Individual agencies will, and should, continue to pursue new capabilities that are within the scope of their mission. But our overall success will depend on an ability to see the complete picture, and on an ability to invest where others do not see it as in their interest to do so. This, we believe, should be the principal focus of the National Science Foundation.