Perspectives on Information Retrieval and Speech

James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA

Abstract. Several years of research have suggested that the accuracy of spoken document retrieval systems is not adversely affected by speech recognition errors. Even with error rates of around 40%, the effectiveness of an IR system falls less than 10%. The paper hypothesizes that this robust behavior is the result of repetition of important words in the text—meaning that losing one or two occurrences is not crippling—and the result of additional related words providing a greater context—meaning that those words will match even if the seemingly critical word is misrecognized. This hypothesis is supported by examples from TREC's SDR track, the TDT evaluation, and some work showing the impact of recognition errors on spoken queries.

1 IR and ASR

Information Retrieval (IR) research encompasses algorithms that process large amounts of unstructured or semi-structured information, though most work has been done with human-generated text. Search engines (such as those on the Web) are a highly visible outgrowth of IR research, where the problem is to present a list of documents (e.g., Web pages) that are likely to be relevant to a user's query. All of the techniques that are needed to find documents, to extract their key concepts, to recognize and avoid "spam" words, to rank the likely matches, and to elicit feedback from the user, etc., are aspects of IR.[1]

In the last several years, automatic speech recognition (ASR) systems have become commercially available. Their introduction extends the realm of IR to include not just text but also audio documents and queries. That is, a passage of speech can be processed by an ASR "speech to text" system and then traditional text-based IR techniques can be employed to help the user locate speeches of

[1] The field of Information Retrieval naturally includes myriad other research issues, ranging from formal modeling of the problem to engineering systems that work across languages, from document clustering to multi-document summarization, and from classification to question answering. In this paper I will focus on search engine technology, though many of the ideas and directions apply equally well to other IR research problems.

interest—directly from the audio (rather, automatically generated text) and not from a human-generated transcript.

Unfortunately, ASR systems are imperfect. Although they do well enough to be marketed commercially, they still make large numbers of errors in the recognition process. The result is an automatically generated transcript where numerous words have been interchanged with other words that (usually) sound vaguely similar to the original. This corruption of the words is potentially a significant problem for IR systems that rely primarily on word matching to find relevant documents.

Even before ASR systems were available as commercial products, IR and ASR researchers had begun work toward exploring interactions between the two fields. The results were somewhat surprising: even tremendous numbers of speech recognition errors in a spoken document had little impact on retrieval effectiveness. Yes, there were theoretical situations where even a single error could make a system fail completely. However, those problems did not seem to crop up in experimental settings.

In the rest of this paper, I will discuss my feelings about why ASR errors have not been a major problem for IR to date. In Section 3 I will support those ideas by reviewing several research papers that explored those ideas. Then, in Section 4 I will show where IR systems begin to break down in the presence of ASR errors, allowing me to claim in Section 5 that significant and interesting open problems remain. I will conclude in Section 6 by discussing several opportunities and additional problems that may arise in the future.

2 Why ASR Is Not a Problem for IR

Why might recognition errors cause problems for information retrieval? Since IR techniques rely fundamentally on matching words and phrases in documents to corresponding items in the query, any process that corrupts the document may cause problems. Indeed, if a critical query term were corrupted in the document, there seems no chance of successful retrieval at all!

I believe this issue is something of a red herring. It is indeed a possibility, but it is not likely to be an issue often. The reason is that documents contain numerous words and it is unlikely that all of them will be corrupted—even a 50% word error rate means that at least half of the words are correct. Other occurrences of that "critical query term" may be properly recognized—and if it is a critical word in the documents, it is almost a given that it will be repeated. Further, even if by some chance all occurrences of that critical term were misrecognized, the documents include numerous other words that provide a context for that word, and their appearance is likely to compensate for the missing word (for example, the word *earthquake* might have words such as *quake*, *tremor*, and *aftershock* providing added context). This theory does, of course, require that queries include multiple words so that context is available.

The reason ASR errors are fairly easily tolerated is the same reason that word sense disambiguation is rarely a problem in information retrieval. There is

an idea that one way to improve the effectiveness of IR systems is to develop a technique for automatically disambiguating the user's query: does *bank* refer to money or rivers, does *fly* refer to airplanes, insects, or trousers? In theory, if a system could automatically figure out the sense of the word intended by the user, retrieval false alarms could be reduced. [11]

However, an ambiguous query word can be made clear by the addition of one or two additional words just as easily: *bank loan*, *river bank*, *fly a plane*, or *buzzing fly*. In all those cases a single additional content word means that the problem of an ambiguous query has essentially been removed.

Now consider what would happen if that single word were surrounded by an entire document that talked about the same topic. A document that talks about flying a plane will include various forms of the root *fly*, as well as things that are flown, things one does to prepare to fly, while flying, and so on. Most documents contain a huge amount of context that support the single word, meaning that that term is much less important than it might seem initially.

To be sure, ambiguity can be a problem (the query *fly* is always ambiguous) and the corruption of that single word in an ASR document might be unrecoverable. However, in the same way that additional words can disambiguate a query, they can also prevent even moderate levels of ASR errors from dropping IR effectiveness too much.

In the next section I will briefly outline a set of experiments from TREC and TDT that support my hypothesis. In Section 4 I will show where ASR errors begin to have an impact on IR effectiveness.

3 Spoken Documents

There have been two major multi-site evaluations that explored the impact of ASR errors on document retrieval and organization. In this section, I outline TREC's and TDT's efforts in that direction, and show how their results are consistent with the hypothesis above.

3.1 TREC 1997 to 2000

From 1997 (TREC-6) through 2000 (TREC-9), the TREC evaluation workshop included a track on "spoken document retrieval" (SDR). [7, 9, 8, 6] The purpose of the track was to explore the impact of ASR errors on document retrieval. The SDR track followed on the heels of the "confusion" track that examined the same question for errors created by OCR (optical character recognition) scanning of documents. [10]

An excellent paper by Garofolo, Auzanne, and Voorhees (2000) summarizes the first three years of the SDR track and comes to conclusions similar to mine. The summary results below are largely due to their analysis. After TREC-9 completed, the conclusion was essentially that SDR is a "solved problem" and that TREC's efforts would be better spent on more challenging problems.

In 1997 (TREC-6) the SDR track was a pilot study to explore how difficult the task would be. A small evaluation corpus of about 50 hours of speech, comprising almost 1500 stories, was used along with 50 "known item" queries. That is, the queries were constructed such that there would be a *single* document known to contain the answer, and such that it was known precisely which document that was (though not by the systems). The reason for the small corpus was that in 1997, recognizing 50 hours of speech was time-consuming, and that was about the limits of technology. The reason for using known item search rather than ranked retrieval is that the former is substantially simpler to assess. The *conclusion* of TREC-6 was that ASR errors caused about a 10% drop in effectiveness (i.e., ability to find that known document at the top of the ranked list). The drop seemed to be consistent, regardless of whether the queries were deemed easy or problematic for ASR errors.

The following year, TREC-7 made the problem more challenging by switching to *ranked* retrieval where there are multiple relevant documents per query. The corpus grew to 87 hours (almost 2900 stories—still very small by IR standards) but only 23 queries were used. Most sites ran their IR systems on a range of ASR outputs, providing a window into the impact of ASR errors (word error rate) on effectiveness. The results showed a clear progressive impact from increasing ASR errors, but only a *small* drop in effectiveness (here, average precision) even when the word error rate climbed to 30–40%.

In TREC-8 (1999), the corpus was made substantially larger so that it was somewhat "reasonable" by IR standards: 550 hours, making up almost 22,000 stories. This time, 50 queries were used in the evaluation. The result: word error rate had minimal impact on average precision. In fact, the results were comparable to the TREC-7 results, even though the corpus was an order of magnitude larger.

TREC's involvement with SDR ended in 2000 with TREC-9. In that case, the corpus was the same as in the previous year, and the same number of queries were used. Again, the ASR errors had minimal impact.

Interestingly, substantially shorter queries were used for TREC-9. For example, here are two forms of the same query:

- SHORT: Name some countries which permit their citizens to commit suicide with medical assistance
- TERSE: assisted suicide

A "short" query means there is much less context within the query, so one might expect the effectiveness to drop. However, the "terse" queries were *more* effective than the "short" queries. This unlikely result may be because of how queries were constructed. In the example listed, the terse query is a phrase that is probably used commonly to describe the issue, whereas the longer query does not include the same important phrase.

Throughout the TREC SDR evaluations, even error rates of about 40% had only a modest impact on IR effectiveness. The length of the recognized speech provided enough repetition of important words, and enough related contextual

words, that the IR retrieval methods could readily compensate for the ASR errors.

3.2 TDT 1998

The Topic Detection and Tracking (TDT) evaluation workshop investigates ways that broadcast news stories can be automatically organized by the events they discuss. [2] That is, all stories about the Oklahoma City bombing should be grouped together, as should stories about a particular earthquake, or a specific political rally. All TDT tasks are carried out on a stream of arriving news stories rather than on a static collection. Because the focus of the work is on broadcast news—i.e., audio streams—speech recognition has always been a key component of TDT.

Tasks in TDT include automatically segmenting the broadcast news stream into topically coherent stories, grouping all stories discussing a particular event together, identifying when a new (previously unseen) event appears in the news, and tracking an event given a handful of on-topic stories. All of these tasks were carried out using both human-generated and ASR-generated transcripts. Note that in contrast to the IR task, "queries" do not exist: most of the tasks require comparing stories to each other, meaning that a "query" is effectively an *entire* document, or even a handful of documents.

The TDT 1998 evaluation used two months worth of news from several sources, approximately 13,000 news stories.[2] The results were:

- For tracking, ASR errors had a modest impact on effectiveness, but not a substantial drop. [13]
- For clustering the stories into topics, the ASR errors had almost no impact on effectiveness.[4]
- When detecting the onset of a new event, ASR errors *did* have a substantially greater impact.[1]

It is not entirely clear why ASR errors had a larger impact on the new event detection. I suspect it is because the task itself is so difficult (effectiveness is generally poor) that ASR errors have a large impact on the ideal parameters for the approach—that is, that better and more training data may find parameters that close the gap.

For two of the TDT tasks, the ASR errors appeared to have little impact. In all three tasks, the effectiveness drop because of ASR errors was very small compared to the overall error rate of the tasks.

4 Spoken Queries

The previous section shows that document retrieval and comparison are fairly robust to ASR errors, even as high as 40% word error rate. Those results support

[2] The TREC-8 and TREC-9 corpus was created from three months of the TDT training data as well as the TDT evaluation data.

my hypothesis that the long documents alleviate the expected problems due to "out of vocabulary" and otherwise misrecognized words.

In this section I look at the other end of the spectrum. Suppose the recognized items were much smaller—e.g., the queries rather than the documents. What, then, would be the impact of ASR errors? According to the hypothesis, the shorter duration of speech will provide less context and redundancy, and ASR errors should have a greater impact on effectiveness.

One group of researchers who investigated this problem [3] considered two experiments. In the first experiment, they recorded 35 TREC queries (topics 101-135). These were quite long queries by modern standards, ranging from 50–60 words, so one might expect that the impact of ASR would not be great. They modified their speech recognition system to produce word error rates of about 25, 33, and 50%. They found that in comparison to an accurate transcription (i.e., the original TREC query), the precision at 30 and at 500 documents dropped about 10% for up to 33% word error rate, and about 15% for the 50% rate. Interestingly, the very top of the ranked list (precision at 5 documents retrieved) saw an *improvement* of 2–4%. These drops in effectiveness are comparable to those seen when documents contained recognition errors.

The same researchers tried another set of substantially shorter queries. Here they created their own queries of three lengths: 2-4, 5-8, and 10-15 content words. The queries were controlled to ensure that there would be a reasonable number of relevant documents retrieved with the accurate transcription of the query (against a corpus of about 25,000 Boston Globe news stories). In this case, the results showed substantial drop in effectiveness, ranging from about 35% worse for 30% word error rate, to 60% worse at 50%. As the queries got slightly longer, the drop in effectiveness became less. Table 1 summarizes the results from their study.

Table 1. Comparison of precision loss for different lengths of queries and different ASR error rates.[3] Caution: ASR error rates are approximate, and the precision is calculated at 30 documents for the long queries and 15 for the rest.

	Word error rate		
	25%	33%	50%
Long (TREC,50-60)	-11%	-9%	-13%
Medium (10-15)	-34%	-41%	-53%
Short (5-8)	-32%	-39%	-57%
Terse (2-4)	-38%	-46%	-61%

These results were verified by Crestani [5] who did some deeper analysis of the long queries. He showed that the ASR degradation was uniform across a range of recall levels (rather than just at a few cutoff points). He also broke the long queries into two groups, longer and shorter than 28 words. He showed that

although both groups degrade in the presence of ASR errors, the longer "long" queries are consistently more accurate than the shorter "long" queries.

Two groups have independently found that the effectiveness of IR systems degrades faster in the presence of ASR errors when the queries are recognized than when the documents are recognized. Further, once queries are less than 30 words, the degradation in effectiveness becomes even more pronounced. These results suggest that there is a minimum number of words necessary for the redundancy and context effects to overcome problems due to ASR errors.

5 Why ASR Is Still an Issue

In Section 3 IR and ASR was described as a solved problem. The previous section shows that this claim is not valid when the recognized item is the query and not the document—at least, not when the query is shorter than about 30 words. For document retrieval, long enough spans of speech can be readily handled.

However, information retrieval is not just about document retrieval. There are other problems in and around IR where ASR is still likely to be a problem. To see where those are likely to be, consider any technology that works on fairly short spans of text. Such a technology, when faced with ASR errors, is unlikely to find enough context and enough redundancy to compensate for the recognition failure. For such technologies, a *single* word incorrectly processed could theoretically have a profound impact.

What does this mean in terms of open problems related to speech within information retrieval systems? Here are several issues that crop up because of the length of the material being used:

- *Spoken questions of short duration.* As shown in Section 4, the drop in effectiveness is large for short spoken queries. How can the ASR be improved for very small snippets of speech? Is it possible for the IR system to guess that the recognition may be bad—because, for example, the query words do not make sense together? Is it just a user interface issue, where people need to be encouraged to talk longer?
- *Message-length documents.* Since short spoken items are the problem, what happens when the documents are substantially shorter? For example, voicemail messages, announcements, and so on.
- *Question answering.* Current technologies to solve the problem of question answering (returning a specific answer to a question rather than just a document) tend to focus on small passages of text that are likely to contain the answer. Finding small passages in the presence of ASR errors may be an issue—and the natural language processing needed to analyze the passages may also fail.
- *User interfaces.* Spoken documents often come grouped together (e.g., a news show with several stories) and need to be broken into segments. How can a user interface properly handle those segments, particularly when the segmentation is likely to contain errors? How can a user "skim" an audio recording to find out whether it is, indeed, relevant? It is possible to skim

text, but audio must be processed linearly. Can a system provide hints to a user to help in this process?

5.1 Can ASR Systems Help ASR/IR?

To address the open problems, it seems useful to consider what information an ASR system might provide rather than just the resulting transcript. There are obvious things that an ASR system can provide to an IR system, but it is not yet clear how the IR system might use them. For example, attempts to use the word lattices or confidence values from a recognition system have yet to provide any benefit. [12] However, the following are things that might have value:

- *Word recognition probabilities.* Although these values have not yet been able to improve IR substantially, they are likely to be more useful for the tasks that focus on smaller portions of text.
- *Word recognition lattices.* Similarly, having access to alternate recognition possibilities may be helpful in tasks such as question answering.
- *Speaker identification.* Knowing the speaker, or some information about the speaker, or even just when the speaker changes, may prove useful for systems that need to identify material from a particular source.
- *Prosody.* It's not clear whether or what types of prosodic information can be reliably extracted from speech, but the "color" it provides to the text may be critical in some applications. For example, consider the value added if it were possible for a system to use prosodic information to isolate sarcasm.
- *Language models.* ASR systems generally contain a language model that provides word sequence probabilities. Some information retrieval systems also rely on language models for retrieval. It is likely that some information could be usefully shared across the two types of models.

5.2 Focusing the System

Another possibility for improving the capabilities of an IR system on spoken data is to focus its attention on a specific person or domain.

- *Voice-id.* Knowing who the speaker is would allow an IR system to bring a user-specific language or topic model to bear when processing the query. Such a model might clarify terms that would otherwise be ambiguous, providing a context for the *likely* meaning of the terms.
- *Domain knowledge.* Both ASR and IR could benefit from collection-specific models of language use: knowing that the user is querying a medical database should make it possible to do better recognition, and to handle the retrieval better. If the system is targeted to a specific task (e.g., a tourist information kiosk), it could adapt both parts of the system, too.
- *User interaction.* Recognizing how the user is interacting with the system might provide useful feedback. For example, if the user is getting annoyed (e.g., more stress in the voice), the IR system might be able to try alternative ways of processing the query, or might ask the user for clarification, and so on.

6 Opportunities and Problems

The success of spoken document retrieval allows the consideration of a much wider range of applications. In this final section, I briefly mention some of the interesting directions that research could go (by no means an exhaustive list!). In doing that, I am also highlighting the problems that ASR/IR systems currently have and that need to be surmounted.

6.1 New Possibilities

Readily available speech recognition has made some things possible that had rarely been considered before. Retrieval of spoken documents is the most obvious possibility. But a range of new applications are now available:

- Coping with very short queries;
- User interface design to encourage longer spoken queries;
- Helping a user decide of a spoken document is a match to his or her query;
- Summarizing speech;
- Adapting ASR language models to the collection being searched;
- Coping with collections that contain a mixture of ASR and clean text documents (typically the latter are more highly ranked);
- Dealing with other types of spoken documents such as dialogues, meetings, lectures, classes, etc.
- Contexts where the word error rate is well over 50%
- Text interfaces to speech-only systems (e.g., voicemail)

6.2 Has ASR/IR Helped IR?

Earlier I mentioned some ways that ASR might be able to help IR, by conveying more information from the recognition process to the IR system. The process of integrating IR and ASR has helped non-ASR retrieval as well, though in subtler ways. ASR techniques have resulted in more robust weighting schemes, in techniques that should be able to cope with misspellings, and with the idea that it makes sense to expand a *document* rather than a query.

7 Conclusion

In this paper I have claimed that for classic information retrieval tasks such as document retrieval, speech recognition errors generally are either inconsequential or can be dealt with using simple techniques. I have softened that somewhat with the acknowledgment that recognition errors are and will be an issue for any language-based technology that looks primarily at small spans of text. When there are only a few words available, there is no opportunity for repetition and context to compensate for errors.

I believe that the interesting challenges ahead for speech applications and information retrieval are suggested by a broader use of spoken documents. What

does it mean to retrieve a meeting? How can IR systems cope with the misstatements, corrections, and disfluencies that are common in less formal speech? Can IR systems benefit from recognition systems that "clean up" the speech as it is transcribed—e.g., adding punctuation, etc?

Finally, and perhaps the biggest challenge, is how user interfaces can be designed to make it possible for people to sift through spoken documents as rapidly as they can pick through clean text? The inherent linearity of speech prevents rapid scanning, and current recognition error rates make it possible to retrieve accurately, but do not necessarily allow a human to get a good "gist" of a retrieved item.

References

1. J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, final report. Available at http://www.clsp.jhu.edu/ws99/tdt, 1999.
2. James Allan, editor. *Topic Detection and Tracking: Event-based News Organization*. Kluwer Academic Publishers, 2001.
3. J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo. Experiments in spoken queries for document retrieval. In *Proceedings of Eurospeech*, volume 3, pages 1323–1326, 1997.
4. Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D. Brown a nd Tom Pierce, and Xin Liu. CMU report on TDT-2: Segmentation, detection and tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120. Morgan Kauffman Publishers, 1999.
5. F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of the 2000 Flexible Query Answering Systems Conference*, pages 267–281, 2000.
6. J. Garofolo, J. Lard, and E. Voorhees. 2000 TREC-9 spoken document retrieval track, 2001. Powerpoint presentation at http://trec.nist.gov.
7. J. Garofolo, E. Voorhees, V. Stanford, and K. Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of TREC-6 (1997)*, pages 83–92, 1998. NIST special publication 500-240.
8. J.S. Garofolo, C.G.P. Auzanne, and E.M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of TREC-8 (1999)*, 2000. NIST special publication 500-246.
9. J.S. Garofolo, E.M. Voorhees, C.G.P. Auzanne, V.M. Stanford, and B.A. Lund. 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of TREC-7 (1998)*, pages 79–89, 1998. NIST special publication 500-242.
10. P. Kantor and E. Voorhes. Report on the TREC-5 confusion track. In *Online proceedings of TREC-5 (1996)*, pages 65–74, 1997. NIST special publication 500-238.
11. R. Krovetz. *Word Sense Disambiguation for Large Text Databases*. PhD thesis, University of Massachusetts, 1995.
12. A. Singhal, J. Choi, D. Hindle, and F. Pereira. AT&T at TREC-6: SDR track. In *Proceedings of TREC-6 (1997)*, pages 227–232, 1998. NIST special publication 500-240.

13. P. van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *Proceedings of the DARPA Broadcast News Workshop*, pages 77–80. Morgan Kauffman Publishers, 1999.