Query By Babbling: A Research Agenda

Douglas W. Oard
College of Information Studies and UMIACS, University of Maryland, College Park, MD, USA
Human Language Technology Center of Excellence, The Johns Hopkins University, Baltimore, MD, USA
oard@umd.edu

ABSTRACT

The Spoken Web, an interconnected collection of spoken content accessed through audio-only cell phones, holds the promise of transforming information access for users in developing regions. The scale of the Spoken Web is, however, limited because current speech retrieval technology is only affordably deployable for a handful of languages. This paper proposes rethinking the conventional keyword query paradigm to instead develop systems that support a longer, richer, and more fluid interaction style that is better suited to both the affordances of spoken interaction and to the limitations of current speech technology.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Measurement

Keywords

Spoken Web, speech retrieval

1. INTRODUCTION

Web search engines have come to define the way people search for information in the minds of both searchers and search engine designers. When faced with a new type of search task it is therefore natural for designers to first think of evolutionary adaptations of the Web search interaction paradigm. After all, early newspapers looked like the pamphlets that had preceded them, early automobiles looked like their horse-drawn predecessors. and Edison's first light bulbs were designed to look like earlier kerosene lamps [5]. Design often proceeds by analogy, and thus it is natural, and perhaps inevitable, to start from what we know. Such has been the case with search for the Spoken Web, an interconnected set of spoken content that users, often users with low literacy, access using ordinary audio-only mobile phones. Our thesis in this paper is that by radically reconceptualizing the interaction paradigm, we can generate new opportunities, although of course we do so at the cost of introducing new challenges as well. We therefore seek to lay out a new interaction paradigm for Spoken Web search, which we call "query by babbling."

Websters Dictionary defines to babble as "to talk enthusiastically or excessively." Indeed, that is precisely what we advocate. Web queries are short, and short queries can indeed work well when the query is text and the content being searched is also text. Indeed, the Web is so large that, even though writers may use language differently, someone will often have used the same words as the searcher to express the same meaning. Said another way, when "early precision" is the goal, text is sufficiently well standardized to permit a reasonable degree of search quality. Such is not the case for speech, however, where people can, and do, speak the same words in quite different ways.

One conventional approach is to convert both query and content to text, and then proceed as in text search (albeit with lower search quality due to speech recognition errors). Such an approach is only practical for a handful of languages, however, since the present technology for building speech-to-text systems is both costly and language-specific. In this paper, we propose an alternative, one based on phonetic matching between query and content. Such an approach is inherently error prone, however, and thus we need more hints from the searcher about what they are looking for. Hence our interest in babbling; babbling is something that some people are good at (and that hopefully others can become good at), and it gives us more chances to correctly infer what it is the searcher is looking for.

The remainder of this paper is organized as follows. In Section 2 we review the work to date on searching the Spoken Web, on indexing speech, and on query formulation. We then present a notional design for a query-by-babbling search engine in Section 3, followed by some remarks on evaluation in Section 4. Section 5 concludes the paper with a few comments on the potential broader implications of this work.

2. BACKGROUND

The first prerequisite for searching speech is that we must have a speech collection that is larger (or at least potentially larger) than could reasonably simply be listened to. Moreover, for speech-to-speech phonetic matching to be the best approach, this speech must be in languages for which good

IKM4DR '12 Maui, Hawaii USA

speech-to-text systems are not presently available, and for which spoken queries can plausibly be formed in the same language as the content. The Spoken Web is one such application, so we begin by introducing the Spoken Web and summarizing the work to date on search in that context. Phonetic indexing has a long heritage in speech retrieval, so we next briefly review that work with a focus on recent developments in which spoken content is matched phonetically without resorting to a text representation at any stage in the process. Finally, query by babbling requires that we effectively infer the searcher's intent from long queries. We therefore conclude this section with a review of what is known about the propensity of searchers to create long queries, and about the ability of search engines to effectively leverage long queries to improve search quality.

2.1 The Spoken Web

The Spoken Web is a spoken analogue to the World Wide Web that is designed to provide access to spoken content using inexpensive audio-only cell phones. Its goal is to transform information sharing and information access among low-literacy users in developing regions. Like people everywhere, such users need many types of timely and reliable information. Among these are information about the prices and availability of goods and services, information about natural and anthropogenic events that may affect their lives, and information support for learning. In a world that has recently been transformed from information scarcity to information abundance for much of the world's population, many low-literacy users remain mired in an earlier era of information scarcity. This happens not because they lack a connection to the Internet, but rather because the Internet currently lacks adequate support for the type of connection that they have: about half the population of the planet, more than three billion people, are presently connected to the Internet only through the two-way audio capabilities of a mobile phone.

It may initially seem strange to see an audio-only mobile phone described as an Internet access device. A little reflection should, however, indicate that its not so strange after all. Blind users rely on screen readers to describe Web pages that were created originally for the sighted. We can have our email read us over the phone. And Siri will do its best to answer our spoken questions with spoken answers. The Internet is already speaking to us, and listening to us, and mobile phones provide natural ways of extending connectivity that we find convenient in some circumstances. If, that is, you happen to speak one of the two dozen or so languages for which sophisticated speech processing technology is available. If not, then your mobile phone would be no better connected to the Internet than a rock would be.

One ambitious effort to bridge this particular digital divide is the Spoken Web [1]. The key idea that underlies the Spoken Web is that the content is "born spoken." This avoids the need to render text as speech and speech as text—the Spoken Web is all speech, all the time. In a sense, it is a parallel Web; interconnection between the World Wide Web and the Spoken Web have yet to appear, despite the obvious potential. This isolation offers a degree of insulation from the dominance of economically powerful languages such as English and Chinese, which together account for about half of all users, and about two-thirds of all content, on the World Wide Web.

This ability to focus equally on economically disadvantaged languages comes with an unprecedented challenge, however: if the Spoken Web is to be scalable in such languages, search technology that does not require the expensive development of language-specific speech processing will be needed. Moreover, this search technology will need to be usable by low-literacy users, for whom many of the interface metaphors that we take for granted (e.g., hierarchical menu navigation) may not be easily mastered [17].

One way of addressing this challenge is to impose some structure on the information space. On the Spoken Web we know what phone the user is calling from, and using that information we can infer some social relationships can provide some leverage [24]. Faceted search based on clustering or classification also offers some potential for supporting effective refinement of search results [10]. But there is simply no scalable substitute for content-based search as a way of getting started [2, 9]. With this goal in mind, the MediaEval shared task evaluation venue created an evaluation design for content-based search of the Spoken Web in four languages in 2011 [22]. In its first two years, the MediaEval Spoken Web task (which continues in 2012) has focused on short queries, but if query by babbling were to be shown to be feasible, MediaEval could offer a natural evaluation venue.

2.2 Phonetic Indexing of Speech

Speech-to-text systems, conventionally called Automatic Speech Recognition (ASR), work by first characterizing the acoustic characteristics of the speech, then inferring higher-level phonetic units from the characterized acoustic sequence, and finally inferring the spoken words from the sequence of phonetic units. One limitation of ASR systems is that they can only recognize known words, those contained in their pronunciation lexicon. When words that are outside an ASR system's vocabulary are spoken, the system must guess some plausible sequence of (often short) known words.

Because employing rare words in a query can be an effective way of focusing a search, there has recently been considerable work on efficient phonetic search for unknown words, a problem known as Spoken Term Detection (STD) [16]. The key idea in STD is to index sequences of phonetic units in a way that permits rapid detection of regions where some specific phonetic sequence might have been spoken, and then to use pronunciation rules at query time to generate plausible phonetic sequences for a desired query term. Although it is possible to build a complete speech retrieval system using such an approach, full ASR systems typically achieve greater accuracy (for known words) because words offer greater potential for modeling context than do phonetic sequences [18]. Because the pronunciation lexicon is known at query time, it is also possible to search known words in a word index built using ASR, using STD only for the unknown query terms, and then to combine the results from the two approaches [20].

Unfortunately, these approaches do not scale well to languages for which the investment necessary to create a well engineered combination of phonetic recognition and a pronunciation lexicon can be justified. Recent work has, however, shown some promise from directly matching phonetic sequences. The key ideas are that relatively long repeated phonetic sequences can be efficiently recognized, and that long phonetic sequences tend to be associated with relatively rare (and hence rather specific) words or phrases. Interestingly, there is now evidence that different speakers use similar phonetic sequences sufficiently often to support detection of different speakers who are discussing the same topics [11]. This is the first of two key technologies that might potentially support query by babbling in languages for which ASR and STD systems are not available.

2.3 Effectively Employing Long Queries

Studies of the behavior of Web searchers invariably indicate that long queries are uncommon. Belkin et al provocatively suggest that this observed behavior may actually be caused by the search engines themselves. To explore this question, they ran a user study in which one condition included a typically small query box and the contrastive condition included a much large text area into which the user could type their query. Intriguingly, users typed substantially longer queries in the larger text area [6]. Although this doesn't tell us that we will be able to get users to babble at length about what they are looking for, it does suggest that short queries are not simply a fact of life.

Indeed, users often pose more than one query in the course of a search session, and information retrieval researchers sometimes reconstruct ersatz "long queries" by segmenting a user's activity into sessions that plausibly reflect a single extended information seeking episode [12]. We might interpret this as indicating some proclivity on the part of users to engage in extended and expressive information seeking episodes of the type we envision in query by babbling.

The key insight behind query by babbling is that long queries provide more opportunities to get the match right. Of course, they also provide more opportunities to get the match wrong. So the key is to get more of the former and less of the later. Recognizing and emphasizing the right terms in long queries has been shown to be important [7], and automated techniques for query shortening by in some way selecting the most useful terms have been shown to be helpful [3, 13]. Lease points out that term selection is an extreme case of term reweighting, arguing (and showing experimentally) that reweighting can yield even better results [14], although perhaps at some cost in efficiency. This line of work on term selection and/or reweighting is the second of the two key technologies that could make query by babbling sufficiently effective to be useful as one basis for searching the Spoken Web.

3. DESIGN

The key idea in query by babbling is to reconceptualize the interaction design of a search engine; this inarguably has substantial consequences for system design, but the first key is to get the interaction design right. We want two things from the searcher: unbounded length (in the same sense that sessions can contain an unbounded number of queries) and rich variety, both in lexical choice and in speaking style. The widely used think-aloud method in user studies [25] offers an intriguing analogue from which to begin an interaction design. Experience shows that study participants can learn to speak in a stream-of-consciousness style, expressing their thought process as they work.

Of course, we don't want our participants to speak endlessly; at some point we want to give them some search results. Essentially, we envision a mixed-initiative searcher-system dialog [19] in which the system can interrupt the searcher (with search results) and the searcher can interrupt they system (to return to babbling in the hope of subsequently generating better results). Both babbling and this "barge in" style for managing turn taking may not come naturally, so we will also need some form of embedded training to help searchers learn to use the system effectively.

A key technical challenge for the system design will be to determine when the system should barge in with some search results. If the system barges in too early, search quality can be adversely affected; too late and we risk the searcher abandoning their search. To find the Goldilocks "just right" times to interrupt, we will need some form of incremental algorithm [23]. Initially, the system should focus on both relevance and diversity; after the first results have been presented, avoiding redundancy with previously presented results will become important [8]. This points up a tension, however, since sometimes users wish to look broadly at what's available and then at some point to return to "re-find" some previous search result. Some way to fluidly "rewind the tape" will therefore likely be needed.

Much also remains to be done with phonetic matching and term selection and/or reweighting. To date, direct matching of long phonetic sequences has been demonstrated only in rather restricted experimental settings (notably, in English, and with a "topically clumpy" test collection in which people talk about a limited range of predefined topics). Considerable work needs to be done to determine whether the salient phonetic differentiations in specific languages are adequately captured by existing phonetic recognizers, the degree to which we can leverage long-term use of the same mobile phone by the same user to perform unsupervised adaptation to the way a specific searcher (or content producer) speaks [26], whether "query-talk" will match the content to be searched adequately well, and how well these techniques will work with richer and more representative collections.

4. EVALUATION

We have to answer two basic questions. First, we need to know whether we can build phonetic matching systems that do what we need them to do. This is the role of intrinsic evaluation. Moreover, we need initially need *formative* intrinsic evaluation, focused on learning how best to perform direct phonetic matching. Once we have what appear to be workable techniques, we can undertake formative extrinsic evaluation, evaluation on representative end-to-end tasks with representative users.

4.1 Formative Intrinsic Evaluation

For intrinsic evaluation, we need some collection of people speaking naturally in some language different from the one that we had in mind when creating out phonetic inventory. Moreover, we need to know at least some instances in which different people are speaking about the same things, and we would prefer to know this by finding these instances in naturally occurring speech rather than by prompting people to talk about specific things. This is a tall order, but fortunately two such collections do already exist, having been created for evaluation of speech retrieval. The Topic Detection and Tracking evaluation created a collection of news broadcasts in English, Chinese and Arabic that are annotated for mentions of a consistent set of topics [27]. News broadcasters speak in a rather stylized manner, however, and we would prefer to work with large quantities of somewhat more natural speech. The Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track created a collection that meets that criterion, containing oral history interviews in English and Czech [21]. One possible evaluation design would be to present a segment from one Czech interview as a query, measuring the effectiveness of the system at ranking passages that address the same topic that are found in other Czech interviews. Phonetic recognizers are easily available for other languages, such as English and German. Similar experiments could also be conducted using the English interviews (perhaps using easily available French or Spanish phonetic recognizers).

4.2 Formative Extrinsic Evaluation

Looking back to the early days of information retrieval, Lewis formulated the Perfect Query Paradox to illustrate that in information retrieval the key is not so much the expressive power of the query language as the user's ability to formulate effective queries [15]. The argument goes something like this. It is easy to show the existence of a Boolean query that retrieves any desired set of documents, and moreover that very often such a query will retrieve no other documents. This so-called "perfect query" is formed by taking the disjunction (i.e., the OR) over the perfect queries for each document. The perfect query for an individual document is the conjunction (i.e., the AND) over the words in the document, conjoined with the negation of the disjunction over the words in the language that are not in the document. The perfect query retrieves undesired documents only when two documents in the collection contain exactly the same words but have different meanings, which simply doesn't happen very often in nature for documents of any substantial length. The paradox is that despite the existence of the perfect query, people continue to try to develop "better" information retrieval systems. Surely no system can be better than perfect.

The paradox is resolved by the recognition that regardless of how good a system may be in some abstract sense, what counts is the searcher's ability to harness that power to find what they are looking for. That's precisely where Boolean queries run into difficulty: ordinary people often have great difficulty formulating effective Boolean queries. The Perfect Query Paradox reminds us that no matter how promising an intrinsic evaluation may be, intrinsic evaluation can tell us only how well we have done what we set out to do. To find out if we have done the right thing, we need extrinsic evaluation.

For this, we need to get some people to babble a bit and see what happens. They need not babble into a mobile phone right from the start; we can start with laboratory experiments that produce some canned data that several research teams can use. We might start with the CLEF CL-SR test collections, since we will have characterized the effectiveness of our matching algorithms on those collections, but ultimately we will want to work with Spoken Web content. The MediaEval[1] Spoken Web task offers an attractive venue for such an evaluation. Perhaps a pilot study might be conducted in 2013 simply by adding a few babbling queries.

5. CONCLUSION

Although the Spoken Web provides the motivating application for query by babbling, the technique might find practical application in several other scenarios where open-domain search is required using only an audio channel. Hands-free Web search while driving might be one such example [4], and serving the needs of blind users might be another. Moreover, the Spoken Web will likely not remain isolated from the World Wide Web for long, and techniques initially developed in one setting will surely be appropriated by users and developers in was that we simply can't presently anticipate. In the mean time, the Spoken Web provides an important driving application for development of this new technology.

6. ACKNOWLEDGMENTS

The author is grateful to Nitendra Rajput for proposing search of the Spoken Web as a challenge problem, to Ken Church for enlightening discussions about the potential for zero-resource speech retrieval using direct phonetic matching, and to Ben Van Durme for suggesting long-term speaker adaptation.

7. REFERENCES

[1] S. K. Agarwal, A. Jain, A. Kumar, A. A. Nanavati, and N. Rajput. The spoken web: A web for the underprivileged. *SIGWEB Newsletter*, pages 1:1–1:9, June 2010.

[2] J. Ajmera, A. Joshi, S. Mukherjea, N. Rajput, S. Sahay, M. Srivastava, and K. Srivastava. Two-stream indexing for spoken web search. In *Proceedings of the 20th International World Wide Web Conference, Companion Volume*, pages 503–512, Mar. 2011.

[3] N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 571–578, July 2010.

[4] A. Baron and P. Green. Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review. Technical Report UMTRI-2006-5, University of Michigan Transportation Research Institute, Ann Arbor, MI, Feb. 2006.

[5] G. Basalla. *The Evolution of Technology*. Cambridge University Press, Cambridge, UK, 1988.

[6] N. Belkin, C. Cool, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, and X.-J. Yuan. Query length in interactive information retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 205–212, July 2003.

[7] C. Buckley and D. Harman. Reliable information access final workshop report. Technical report, ARDA Northeast Regional Research Center, Bedford, MA, Jan. 2004.

[8] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Aug. 1998.

[9] K. A. Dhanesha, N. Rajput, and K. Srivastava. User driven audio content navigation for spoken web. In

[1] http://www.multimediaeval.org/

Proceedings of the International Conference on Multimedia, pages 1071–1074, 2010.

[10] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava. Faceted search and browsing of audio content on the spoken web. In *Proceedings of the 19th Annual ACM International Conference on Information and Knowledge Management*, pages 1029–1037, Oct. 2010.

[11] M. Dresde, A. Jansen, G. Coppersmith, and K. Church. NLP on spoken documents without ASR. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 460–470, Oct. 2010.

[12] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 699–708, 2008.

[13] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 564–571, July 2009.

[14] M. Lease. *Beyond Keywords: Finding Information More Accurately and Easily Using Natural Language.* Ph.D. in Computer Science, Brown University, Providence, RI, 2010. http://cs.brown.edu/research/pubs/theses/phd/2010/lease.pdf.

[15] D. D. Lewis. *Representation and Learning in Information Retrieval.* Ph.D. in Computer Science, University of Massachusetts, Amherst, MA, 1992. http://ciir.cs.umass.edu/pubfiles/UM-CS-1991-093.pdf.

[16] J. Mamou, B. Ramabhadran, and O. Siohan. Vocabulary independent spoken term detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 615–622, 2007.

[17] I. Medhi, S. Patnaik, E. Brunskill, S. N. Gautama, W. Thies, and K. Toyama. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction*, 18(1):2:1–2:28, 2011.

[18] K. Ng. *Subword-Based Approaches for Spoken Document Retrieval.* Ph.D. in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2000. http://dspace.mit.edu/bitstream/handle/1721.1/16737/45156861.pdf?sequence=1.

[19] D. G. Novick and S. Sutton. What is mixed initiative interaction? In *AAAI Spring Symposium on Computational Models for Mixed-Initiative Interactions*, pages 114–116, Mar. 1997.

[20] J. S. Olsson and D. W. Oard. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 91–98, 2009.

[21] P. Pecina, P. Hoffmannova, G. Jones, Y. Zhang, and D. Oard. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum*, volume 5152 of *Lecture Notes in Computer Science*, pages 674–686. Springer, 2008.

[22] N. Rajput and F. Metze. Spoken web search. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, Oct. 2011. http://ceur-ws.org/Vol-807/Rajput_SWS_me11overview.pdf.

[23] A. Rostamizadeh, A. Agarwal, and P. Bartlett. Learning with missing features. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 635–642, July 2011.

[24] S. Sahay, N. Rajput, and N. Pansare. Social ranking for spoken web search. In *Proceedings of the 20th Annual ACM International Conference on Information and Knowledge Management*, pages 1835–1840, Oct. 2011.

[25] M. W. van Someren, Y. F. Bernard, and J. A. Sandberg. *The Think-Aloud Method: A Practical Guide to Modeling Cognitive Processes.* Academic Press, London, 1994.

[26] R. Wallace, K. Thambiratnam, and F. Seide. Unsupervised speaker adaptation for telephone call transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4393–4396, Apr. 2009.

[27] C. L. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, May 2000.