Retrieval Effectiveness of Written and Spoken Queries: an Experimental Evaluation

Heather Du and Fabio Crestani

Dept. of Computer and Information Sciences
University of Strathclyde
UK G1 1XH
{heather,fabioc}@cis.strath.ac.uk

Abstract. With the fast growing speech technologies, the world is emerging to a new speech era. Speech recognition has now become a practical technology for real world applications. While some work has been done to facilitate retrieving information in speech format using textual queries, the characteristics of speech as a way to express an information need has not been extensively studied. If one compares written versus spoken queries, it is intuitive to think that users would issue longer spoken queries than written ones, due to the ease of speech. Is this in fact the case in reality? Also, if this is the case, would longer spoken queries be more effective in helping retrieving relevant document than written ones? This paper presents some new findings derived from an experimental study to test these intuitions.

1 Introduction and Motivations

At long last, speech is becoming an important interface between human being and machine. Computer systems, whether fixed or mobile, wired or wireless, increasingly offer users the opportunity to interact with information through speech. The conventional means of information seeking using textual queries is becoming more difficult to satisfy the desire for information access of a mobile user. Accessing information using textual queries does not work well for users in many situations, such as when users are moving around, with their hands or eyes occupied in something else, or interacting with another person. For those with visual impairment such as blindness or difficulty in seeing words in ordinary newsprint, not to mention those with limited literacy skills, speech would be the only means to satisfy their information needs. In all these cases, given the advancement of speech technology, speech enabled interface has come to the lime light of today's information retrieval (IR) research community, with the promise of enabling users to access information solely via voice.

The transformation of user's information needs into a search expression, or query is known as query formulation. It is widely regarded as one of the most challenging activities in information seeking [1]. Research on spoken query formulation and use for information access is denoted as spoken query processing (SQP).

From 1997 (TREC-6) to 2000 (TREC-9), TREC (Text Retrieve Conference) evaluation workshop included a track on spoken document retrieval (SDR) to explore the impact of automatic speech recognition (ASR) errors on document retrieval. The conclusion draw from this three years of SDR track is that SDR is a "solved problem" [2]. This is certainly not the case for SQP.

SQP has been focusing on studying the level of degradation of retrieval performance due to errors in the query terms introduced by the automatic speech recognition system. The effect of the corrupted spoken query transcription has a heavy impact on

the retrieval ranking [3]. Because IR engines try to find documents that contain words that match those in the query, any errors in the query have the potential for derailing the retrieval of relevant documents. Two groups of researchers have investigated this problem by carrying out experimental studies. One group [4] considered two experiments on the effectiveness of SQP. These experiments showed that as the query got slightly longer, the drop in effectiveness of system performance became less. Further analysis of the long queries by the other group showed that [5] the longer “long" queries are consistently more accurate than the shorter “long" queries. In general, these experiments concluded 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 short the degradation in effectiveness becomes even more noticeable [6]. Therefore, it can be claimed that despite the current limitations of the accuracy of ASR software, it can be feasible to use speech as a means of posing questions to an IR system as long as the queries are relatively long. However, the query sets created in these experiments were artificial, being made of queries originally in textual form and dictated. Will spontaneous queries be long? Will people use same words, phrases or sentences when formulating their information needs via voice as typing onto a screen? If not, how different are written queries from spoken ones? What level of retrieval effectiveness should we expect from spontaneous spoken queries? It is a well-known fact that dictated speech is considerably different from spontaneous speech and easier to recognise [7]. It should be expected that spontaneous spoken queries would have higher levels of word error rate (WER) and different kinds of errors. Thus, the conclusions drawn from previous experimentation with spoken queries will not be valid until further empirical work is carried out to clarify the ways in which spontaneous queries differ in length and nature from dictated ones.

In this paper we present the results of an experimental study on the differences between written queries and their counterpart in spoken forms. We also present an evaluation of their respective retrieval performance effectiveness against an IR system. The paper is structured as follows. Section 2 discusses the usefulness of speech as a means of query input. Section 3 describes how we built a collection of spoken and written queries and highlights some of the differences found between the two. This collection of spoken and written queries is the test collection we will employ in our effectiveness study. The results of this study are reported in section 4. Conclusion with some remarks on the potential significance of the study and future directions of work are presented in section 5.

2 Spoken Queries

The advantages of speech as a modality of communication are obvious. It is natural just as people communicate as they normally do; it is rapid: commonly 150-250 word per minutes [8]; it requires no visual attention; it requires no use of hands.

However, ASR systems produce far from perfect transcripts, which means that there is bound to be recognition mistakes at different levels depending on the quality of the ASR systems. Queries are generally much shorter than documents in the form of both text and speech. The shorter duration of spoken queries provides less context and redundancy, and ASR errors have a greater impact on effectiveness of IR systems. Furthermore, input with speech is not always perfect in all situations. Speech is public, potentially disruptive to people nearby and potentially compromising of confidentiality. Speech becomes less useful in noisy environment. The cognitive load im-

posed by speaking must not be ignored. Generally when formulating spoken queries, users are not simply transcribing information but are composing it.

However, despite the unavoidable ASR errors, research shows that the classical IR techniques are quite robust to considerably high level of WER (about up to 40%), in particular for longer queries [9]. In addition, it has long been proved that voice is a richer media than written text [11]. It has more cues including voice inflection, pitch, and tone. Research shows that there exists a direct relationship between acoustic stress and information content identified by an IR index in spoken sentences since speakers stress the word that can help to convey their messages as expected [10]. People also express themselves more naturally and less formally when speaking compared to writing and are generally more personal. Thus, we would expect, as a result, that spoken queries would be longer in length than written queries. To test this hypothesis, we constructed and carried out an experiment as described in the following section.

3 Qualitative Comparison of Written versus Spoken Queries

Our view is that the best way to assess the difference in query formulation between spoken form and written form is to conduct an experimental analysis with a group of potential users in a setting as close as possible to a real world application. We used a within-subjects experimental design [12].

3.1 Experimental Study

As retrieving information via voice is still relatively in its infancy, it would be difficult to identify participants for our study. We therefore decided to recruit from an accessible group of potential participants who is not new to the subject of Information Retrieval. 7 of our participant members were from the IR research group who have good knowledge of IR to some degree and 5 participants were research students who all have good experience of using search engines within the department of computer and information sciences, but few have prior experience with vocal information retrieval. It is worth to mention that all participants were native English speakers. There would be no language barriers for them to understand and formulate their information needs in English.

The set of topics we used for this experimental study was a subset of 10 topics extracted from TREC topic collection (topics 151–160). Each topic consists of four parts: id, title, description and narrative.

The experiment consisted of two sessions. Each session involved 12 participants, one participant at a time. The 12 participants who took part in the first session also took part in the second session. An experimenter was present throughout each session to answer any questions concerning the process at all times. The experimenter briefed the participants about the experimental procedure and handed out instructions before each session. Each participant was given the same set of 10 topics in text form. These topics were in a predetermined order and each had a unique ID. The tasks were that each participant was asked to form his/her own version for each topic in either written form or spoken form as instructed via a graphic user interface (GUI) on a desktop screen (written in Java). For session 1, each participant was asked to form his/her queries in written form for the first 5 topics and in spoken form for the second 5 topics via the GUI. For session 2, the order was reversed, that was each participant presented his/her queries in spoken form for the first half topic set and in written form for the

second half topic set via the GUI. Each session lasted approximately 3 hours, which gave each participant to finish the tasks within 30 minutes and a maximum of 5 minutes time constraint was also imposed on each topic. During the course of the experiment, the written queries were collected and saved in text format. The spoken ones were recorded using close-talk microphone and saved in audio format in a wav file for each participant automatically. The data collected were used for post-experimental analysis and to test the experimental hypothesis.

3.2 Experimental Results

From this experiment, we have collected 120 written queries and 120 spoken queries that have been manually transcribed. Some of the characteristics of written and spoken queries are reported in Table 1.

This table pictures clearly that the average length of spoken queries is longer than written queries with a ratio rounded at 2.48. This seems to confirm our hypothesis that spoken queries are longer than written ones. After stopwords removal, the average length of spoken queries is reduced from 23.07 to 14.33 with a 38% reduction rate and the average length of written queries is reduced from 9.54 to 7.48 with a reduction rate at 22%. These figures indicate that spoken queries contain more stopwords than written ones. This indication can also be seen from the differentials between the average length and median length for both spoken and written queries. The difference between the numbers of unique terms occurred in the written query set and spoken query set is not great. This is because each participant worked on the same 10 topics and generated a written query and a spoken query for each topic, therefore there are 12 versions of written queries and 12 versions of spoken queries in relation to one topic.

Table 1. Characteristics of written & spoken queries.

Data set	Written queries	Spoken queries
Number of queries	120	120
Unique terms in queries	309	552
Average query length (with stopterms)	9.54	23.07
Average query length (without stopterms)	7.48	14.33
Median query length (without stopterms)	7	11

The average length of written and spoken queries with/without stopwords for each topic is calculated and presented in Fig. 1. In Fig. 1, the scattered points for spoken queries always stay above the ones for written queries, which suggests the spoken queries are lengthier than the written ones. This is a case for every topic persistently. This is exactly what we would expect to see. We know from previous studies that the textual queries untrained users pose to information retrieval systems are short: most queries are three words or less. With some knowledge of IR and high usage of Web search engines, our participants have formulated longer textual queries. This is also typical of trained users. When formulating queries verbally, the ease of speech encouraged participants to speak more words

From the above analysis, we know that spoken queries as a whole are definitely lengthier than written queries. One would argue that people with natural tendency would speak more conversationally which results in lengthy sentences containing a great deal of function words such as prepositions, conjunctions or articles, that have little semantic contents of their own and chiefly indicate grammatical relationships, which have been referred as stopwords in IR community, whereas the written queries

are much terser but mainly contain content words such as nouns, adjectives and verbs, therefore, spoken queries would not contribute much than written queries semantically. However, after we remove the stopwords within both the spoken and written queries and plot the average length of spoken and written queries against their original length in one graph, as shown in Fig. 1, which depicts a very different picture.

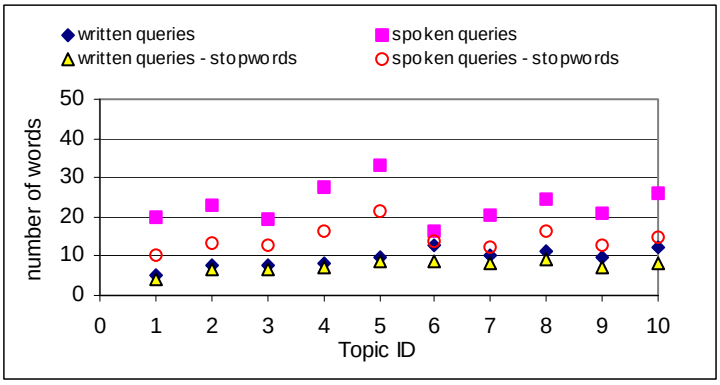

Fig. 1. Average length of queries across topics.

As we can see from this figure, the points for spoken queries are consistently on top of the ones for the written queries; after stopwords removal, each of them are also undoubtedly becoming shorter. Moreover, the points for spoken queries without stopwords stay above the ones for written queries without stopwords consistently across every topic. Statistically, the average spoken query length without stopwords is 14.33 and for written query, that is 7.48, which shows the spoken queries have almost doubled the length of the written ones. This significant improvement in length indicates that the ease of speaking encourages people to express not only more conversationally, but also more semantically. From IR point of view, more search words would improve the retrieval results. Ironically, for SQP, the bane is the very tool that makes it possible: the speech recognition. There are wide range of speech recognition softwares available both for commercial and research purposes. High quality speech recordings might have a recognition error rate of under 10%. The average word error rates (WER) for large-vocabulary speech recognisers are between 20 to 30 percent [13]. Conversational speech, particularly on a telephone, will have error rates in the 30-40% ranges, probably on the high end of that in general. In this case in our experiment, even if at the WER at 50%, it would not cause greater degradations on the spoken queries to make them shorter than written queries. In other word, the spoken information clearly has the potential to be at least as valuable as written material.

We also summarise the length of queries with/without stopwords for all 10 topics across all participants. The average length of queries per participant is presented in Fig. 2.

We could observe from Fig. 2 that it is the same case for every participant that his/her spoken queries are longer than written ones consistently. However, the variations of the length between spoken and written queries for some participants are very small. In fact, after we studied the transcriptions of spoken queries, we observed that the spoken queries generated by a small portion of participants are very much identical to their written ones. The discrepancies of length within written queries are very insignificant and relatively stable. All participants used similar approach to formulate their written queries by specifying only keywords. The experience of using textual search engines influenced the participants' process of query formulations. For most popular textual search engines, the stopwords would be removed from a query before

creating the query representation. Conversely, the length fluctuates rapidly within spoken queries among participants.

We did not run a practice session prior to the experiment such as to give an example of how to formulate a written query and a spoken query for a topic, since we felt this would set up a template for participants to mimic later on during the course of experiment and we would not be able to find out how participants would go about formulating their queries. In this experiment, we observed that 8 out of 12 participants adopted natural language to formulate their queries which were very much like conversational talk and 4 participants stuck to the traditional approach by only speaking keywords and/or broken phrases. They said they did not want to "talk" to the computer as they felt strange and uncomfortable to speak to a machine. This suggests that participants own personalities played a key roll in the query formulation process.

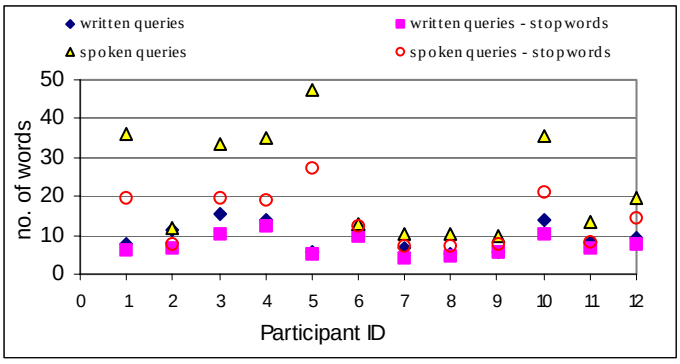

Fig. 2. Average length of queries per participant.

After stopwords removal, the spoken queries are still lengthier than the written ones. Fig. 2 shows a consistency with the result of the previous analysis that people tend to use more function words and content words in speaking than writing. This is true for every participant in our experiment.

A sentence in natural language text is usually composed of nouns, pronouns, articles, verbs, adjectives, adverbs, and connectives. From IR point of view, not all words are equally significant for representing the semantics of a document. Investigating the distribution of different part of speech (POS) in the two forms of queries gives us another opportunity to shed light on the nature of the differences and similarities between spoken and written queries. Fig. 3 shows a comparison of POS between the two query sets. This figure indicates that categorematic words, primarily nouns, verbs and adjectives, i.e. words that are not function words, made up a majority of word types. There are more types of words in spoken queries than written queries. Nouns, adjectives and verbs are frequently used in both written and spoken queries. Nouns have the largest type shares in both query forms and higher percentage in written queries than spoken queries, as nouns are well known to carry more information content and therefore more useful for search purposes. Verbs are the second largest POS in spoken queries and the third largest in written queries thus they seem to play a more important role in spoken than in written queries, whereas adjectives are more common in written queries than in spoken queries. Prepositions and conjunctives are also heavily used in spoken queries. These two POS types are considered stopwords, so they would be automatically removed during the indexing procedure.

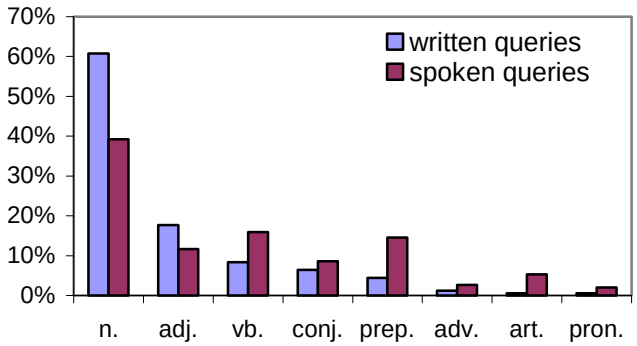

Fig. 3. Percentages of part-of-speech in written & spoken queries.

4 Retrieval Effectiveness of the Written versus Spoken Queries

This section describes the procedure and the results of an experimental analysis into the effectiveness of written versus spoken queries. In this context we assume that the spoken queries have been perfectly transcribed, that is, the speech recognition process is perfect. This is of course a gross simplification, since even well trained ASR systems make recognition mistakes. Nevertheless, we believe this study could provide the upper bound level of performance of an IR system using spoken queries.

4.1 Experimental Procedural

In order to experiment the differences in effectiveness of written and spoken queries, a suitable test environment needs to be devised. Classical IR evaluation methodology [14] suggests that such test environment should consist of the following components:

a) a collection of textual document;
b) a set of written and spoken queries with associated document relevance assessments;
c) IR system;
d) some measures of the IR system effectiveness.

The collection we used is a subset of the collection generated for TREC [15]. The collection is made of the full text of articles of the Wall Street Journal from year 1990 to year 1992. The 120 written and 120 spoken queries collected from above mentioned experiment were used. Since the two sets of queries were generated based on the 10 TREC topics, we could be able to use the corresponding set of relevant documents.

We used the Lemur IR toolkit to implement the retrieval system. Lemur has been developed by the Computer Science Department of the University of Massachusetts and the School of Computer Science at Carnegie Mellon University [16]. It supports indexing of large text collection, the construction of simple language models for documents, queries and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

The main IR effectiveness measures used in our study are the well-known measure of Recall and Precision. Recall is defined as the portion of all the relevant documents in the collection that has been retrieved. Precision is the portion of retrieved documents that is relevant to the query. Once documents are ranked in response to a query according to the retrieval status value (RSV), precision and recall can be easily evalu-

ated. These values are displayed in tables or graphs in which precision is reported for standard levels of recall (from 0.1 to 1.0 with 0.1 increments). In order to measure the effectiveness of the IR system of written and spoken queries, a number of retrieval runs were carried out against different IR models and precision and recall values were evaluated. The results reported in the following graphs were averaged over the entire sets of 120 written queries and 120 spoken queries.

4.2 Results of the Effectiveness of Written and Spoken Queries

We ran these two sets of queries against three models implemented using the Lemur toolkit: a basic TFIDF vector space model, the Okapi, and a language modelling method that used the Kullback-Leibler similarity measure between document and query language models. No relevance feedback methods were used for any of these three models.

The TFIDF vector space model was implemented using standard methods in which each document and each query are represented by term frequency vectors, then the terms in those vectors are weighted using TFIDF weight, and finally the RSV value of each query-document pair is calculated as the sum of their term weights. Fig. 4 depicts the effectiveness of written and spoken queries using the above TFIDF models. Naturally, we would expect the best result should be obtained for the perfect transcript, but the performances obtained for the two query sets are very similar. Like in any scientific experiment, the outcome of an IR experiment is affected by random errors. As the result, we cannot conclude that one is better than the other based on a small performance difference between two query sets. Significance tests are needed to decide whether the performance difference between two query sets is statistically significant. The paired t-test is the most widely used in IR. The general idea behind the tests is: we assume that two techniques being compared are equally good. Under the assumption, we calculate a probability (p-value) that the observed performance difference could occur by chance. The smaller is the p-value, the more significant is the difference between the two techniques. The p value derived from TFIDF retrieval was very big; this indicates that a difference on system performance between spoken and written queries is statistically insignificant.

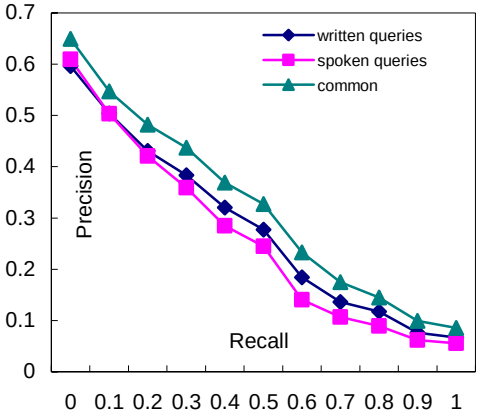

Fig. 4. P/R graph for simple TFIDF model

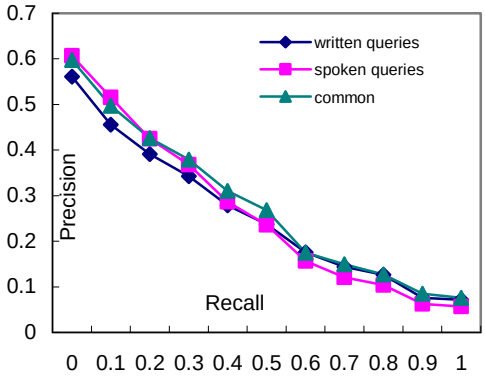

Fig. 5. P/R graph for KL & JM model.

In order to identify which words were responsible for the effectiveness results, we artificially built another query set of 120 queries by using the words appearing in both written queries and spoken queries. The results obtained with this set of queries, shown in Fig. 4, indicate that this query terms set obtained slightly better retrieval performance. This is an indication the important words (those responsible for the retrieval effectiveness) are present in both query sets. Those words that are present only in spoken or written queries are therefore responsible for the decrease in performance.

Only slightly different results were obtained using the language model implemented in Lemur, which is based on the Kullback-Leibler similarity measure and the JM smoothing (KL&JM). Fig. 5 depicts the effectiveness of the three query sets and they all had very similar performance. In fact a t-test shows that the differences are not statistically significant. It seems that the language model is not able to show the difference between the sets of spoken and written queries and the set of queries obtained by only considering the words these two sets have in common.

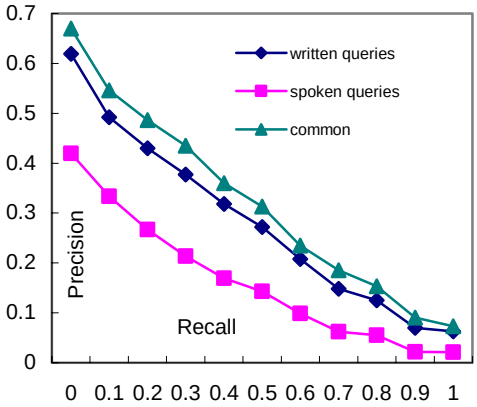

Fig. 6. P/R graph for simple Okapi model

Rather different results are reported in Fig. 6, which shows the effectiveness of written and spoken queries against the Okapi BM25 model, implemented in Lemur using the well-known BM25 formula [17]. In this case we can observe that written queries clearly outperformed spoken ones. The t-test shows that the difference between these two performances is statistically significant. The common query terms set obtained very similar but slightly better performance as the written queries.

It is not clear to us how so different effectiveness measure were obtained for the same query sets using different models. However, it is clear that our hypothesis that spoken queries would produce better retrieval performance just because they were longer does not hold. This warrants further investigation that is still under way and that is only partially reported in the next section.

4.3 Analysis of the Experimental Results

From above Fig. 4 and Fig. 5 reporting the results of the TFIDF and KL&JM models applied to written and spoken queries, we can conclude that these two sets of queries are almost equally effective with respect to retrieval performances. This is contrary to what we expected. From the previous experiment on qualitative comparison between written and spoken queries in terms of their length, we could claim that spoken queries are more useful than written queries because they carry more content words. As far as IR performance is concerned, more content words should lead to a more effective relevant document retrieval. This fact is supported by much past research. So, where have the content words gone during the retrieval process? These two graphs also shows that the performance of the common query terms is very similar to the ones of written and spoken query sets from which it was extracted. This indicates that the words useful for retrieval purposes are those words that appear in both written queries and spoken queries. Lets us look at this result by taking a specific query. A typical user spoken query looks like the following:

> "I want to find document about *Grass Roots* Campaign by *Right Wing Christian Fundamentalist* to enter the political process to further their religious agenda in the *U.S*. I am especially interested in threats to *civil liberties*, government stability and the *U.S. Constitution*, and I'd like to find feature articles, editorial comments, news items and letters to the editor."

Whereas its textual counterpart looks like:

> "*Right wing Christian fundamentalism, grass roots, civil liberties, US Constitution*."

Words present in both queries are reported in italic. The words appearing in the written query are more or less also present in its corresponding spoken query. Other words in spoken query include conjunctions, prepositions and articles that will be removed as stopwords. The parts such as "I want to find document about" and "I am especially interested in" are conversional and contained words that while they will not all be removed as stopwords, will definitely have very low weights (IDF or KL) and therefore would not be useful. Although there are also some nouns in the spoken query, such as "feature articles, editorial comments, news items letters editor" which specify the forms of relevant document, these words are unlikely to appear in relevant documents therefore do not contribute to the RSVs. The vocabulary sizes of these three query sets are shows in Table 2. 71% of words in written queries are in the common words whereas only 40.9% for spoken query words. The ratio of common terms over the total vocabulary sizes of written and spoken queries is 25.9%.

Table 2. Vocabulary size of different query sets.

Written queries	Spoken queries	Common queries terms
309	552	226

Fig. 5 shows the effectiveness of written and spoken queries against Okapi model. Surprisingly, the BM25 formula seems to have a very bad effect on spoken queries. The written queries manage to maintain its performance, whereas the retrieval effectiveness for spoken queries gets much worse than that obtained with the TFIDF and KL&JM models. There is no clear explanation for this phenomenon. A deeper analysis needs to be carried out to study this effect, before any conclusions could be generalized.

5 Conclusions and Future Work

This paper reports an experimental study on the differences between spoken and written queries in qualitative terms and in terms of their effectiveness performance, assuming perfect recognition. This study serves as the basis for a preliminary speech user interface design, to be carried out in the near future. The results show that using speech to formulate one's information needs not only provides a way to express it naturally, but also encourages one to speak more "semantically", i.e. using more content bearing words. This means that we can come to the conclusion that spoken queries as a means of formulating and inputting information needs are utterly feasible.

Information retrieval systems are very sensitive to errors in queries, in particular when these errors are generated by applying ASR to spoken queries [18]. We are fully aware of this potential threat, therefore for future work, we are going to design robust IR models able to deal with this problem. With this goal in mind, we are going first to transcribe the recordings of the spoken queries using ASR software and the identify an IR system which can be used to evaluate the effect of word error rate of spoken queries against written queries on the effectiveness of the retrieval performance. We will then study how the IR system can be made more robust to these errors. One possible way is to use on verbal information contained in speech, like for example prosodic stress, in conjunction with POS tagging to identify the most useful words on which the recognition accuracy of the ASR process should be concentrated.

As a side research, we are carrying out a similar experiment on Mandarin, a language that has a completely different semantic structure from English, to check if the results presented in this paper also hold for other languages. The topics being used for this experimental study are a subset extracted from the TREC-5 Mandarin Track and the participants are all native Mandarin speakers with good experience of using search engines.

Acknowledgments

The authors would like to thank all the participants who were from the Department of Computer and Information Sciences at the University of Strathclyde for their efforts and willingness in taking part in this experiment voluntarily.

References

[1] Cool, D., Park, S., Belkin, N.J., Koenemann, J. and Ng, K.B. Information seeking behaviour in new searching environment. *CoLIS 2*. Copenhagen. (1996)403-416.

[2] J. S. Garofolo, C.G.P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: a success story. In *Proceedings of the TREC Conference*, pages 107-130, Gaithersburg, MD, USA, November 1999.

[3] E. Mittendorf and P. Schauble. Measuring the effects of data corruption on Information Retrieval. In *Proceedings of the Workshop on Speech and Natural Language*, pages 14-27, Pacific Grove, CA, USA, February 1991.

[4] 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.

[5] F.Crestani. Spoken Query Processing for Interactive Information Retrieval. *Data and Knowledge Engineering*, 41(1): 105-124, 2002.

[6] J. Allan: Perspectives on Information Retrieval and Speech. *SIGIR Workshop: Information Retrieval Techniques for Speech Applications* 2001: 1-10.

[7] E. Keller (Ed.), Fundamentals of Speech Synthesis and Speech Recognition, *John Wiley and Sons*, Chichester, UK, 1994.

[8] D. R. Aaronson and E. Colet. Reading paradigms: From lab to cyberspace? Behavior Research Methods, *Instruments and Computers*, 29(2):250–255, 1997.

[9] F. Crestani. Effects of word recognition errors in spoken query processing. In *Proceedings of the IEEE ADL 2000 Conference*, pages 39-47, Washington DC, USA, May 2000.

[10] A. Tombros and F. Crestani. User's perception of relevance of spoken documents. *Journal of the American Society of Information Science*, 51(9):929-939, 2000.

[11] Barbara L. Chalfonte, Robert S. Fish, Robert E. Kraut. Expressive richness: a comparison of speech and text as media for revision. In *proceeding of the SIGCHI conference on Human factors in computing systems: Reaching through technology*. Pages: 21 – 26, 1991.

[12] S. Miller. *Experimental design and statistics.* Routledge, London, UK, second edition, 1984.

[13] Edro J Moreno J-M. Van Thong, Beth Logan. From Multimedia Retrieval to knowledge management. *Computer*, pages 58-66, 2002.

[14] C. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, UK, 1979.

[15] E. Voorhees, D. Harman, Overview of the seventh text retrieval conference (TREC-7), in: Proceedings of the TREC Conference, Gaithersburg, MD, USA, 1998, pp. 1-24.

[16] P. Ogilvie and J. Callan. Experiments using the Lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)* (pp. 103-108). National Institute of Standards and Technology, special publication 500-250.

[17] S.E. Robertson, S. Walker, S. Jones, M.M. HancockBeaulieu, and M. Gatford. Okapi at TREC-3. In D.Harman, editor, *Proc. TREC-3, the 3rd Text Retrieval Conference*, pages 109-127. NIST, 1995.

[18] J. Allan. Knowledge Management and Speech recognition. *Computer*. April 2002, pages 46-47.