Written versus Spoken Queries: a Qualitative and Quantitative Comparative Analysis

Fabio Crestani and Heather Du

Department of Computer and Information Sciences

University of Strathclyde

Glasgow G1 1XH, Scotland, UK

Abstract

This paper reports on an experimental study on the differences between spoken and written queries. A set of written and spontaneous spoken queries are generated by users from written topics. These two sets of queries are compared in qualitative terms and in terms of their retrieval effectiveness. Written and spoken queries are compared in terms of length, duration, and part of speech. In addition, assuming perfect transcription of the spoken queries, written and spoken queries are compared in terms of their aptitude to describe relevant documents. The retrieval effectiveness of spoken and written queries are compared using three different IR models. The results show that using speech to formulate one's information need provides a way to express it more naturally and encourages the formulation of longer queries. Despite that,

longer spoken queries do not seem to significantly improve retrieval effectiveness compared with written queries.

1 Introduction and Motivations

Traditionally, Information Retrieval (IR) has been concerned with retrieving textual documents in response to written queries. Multimedia technologies have recently enabled to describe and index multimedia documents, so that an image, video or even speech can now be retrieved in response to a written query. Only very recently some work has been directed at studying how documents could be retrieved with queries that are not in written form (see the related work section). It is quite surprising that so little work has been devoted to studying systems that retrieve documents by means of spoken queries, given the prominence of speech as a communication medium (see the related work section). The current trend towards remote and mobile access to information is making it necessary to design and develop such systems.

Today, the phone is the most widely adopted communications device anywhere in the world. Mobile phone subscriptions are increasing faster than Internet connection rates. The development of wireless technology enables this huge mobile user community to take advantage of the large amount of information stored in digital repositories and access the information from anywhere and at anytime. Currently, the means of input user's information needs available are very much limited in keypad capability by either keying in or using a stylus on the mobile phone screen. Text-entry rates for the multi-tap method are

commonly 7-15 wpm; with predictive-text facilities this rate roughly doubles (Silfverberg, MacKenzie, & Korhonen, 2000). Key-tapping would therefore allow the entry of a typical 10-word question in 20-40 seconds, with continuous visual attention. Hand-writing with a stylus can double that (Soukore□ & MacKenzie, 1995), but while this would suffice to satisfy some information needs, such input style does not work well for users in many situations, such as when they are moving around, or using their hands or eyes for something else, or interacting with another person. In addition, the availability of screens and keyboards are not useful to users with visual impairment such as blindness or difficulty in seeing words in ordinary newsprint, not to mention those with limited literacy skills. In all those cases, speech enabled interface would let users 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 (Cool, Park, Belkin, Koenemann, & Ng, 1996). This paper is concerned with the study of query formulation and in particular in making a comparative study between written and spoken queries. We assume that a user has an information need and that she wants to use an IR system to retrieve documents that might contain information that satisfy it. Depending on the context and the task in which the user is involved, the query submitted to the IR system could be either written (via a keyboard) or spoken (via a microphone or phone). The research questions we pose ourselves are the following. What are the differences between written and spoken queries in terms

of their retrieval characteristics and performance outcome? How should the unique characteristics of spoken queries be exploited in IR system design and development?

The paper is structured as follows. Section 2 introduces some background and related work. Section 3 reports on an empirical study on the comparison between written and spoken queries. Written and spoken queries are compared in terms of length, duration, part of speech, and retrieval performance. Comments on experience in formulating spoken queries by the study participants are also reported. In the last section conclusions are drawn, as well as a projection for directions of future work.

2 Written Queries versus Spoken Queries

In this section we provide some background on the study reported in this paper and also place the study in relation to related work.

2.1 Background

IR has been dealing with written queries for the whole of its history. Written queries can be in the form of a Boolean statement, involving keywords and Boolean operators, or can be a natural language statement. Because of the way current IR systems work, the query is transformed into a representation that enables quick comparison with document representation. In other words, a natural language expression describing an information need is reduced to a ”bag

of keywords". While this process is relatively easy for written queries, it is much more complex and error-prone for spoken queries, since speech needs to be reduced to text to be processed by the IR system. Automatic speech recognition (ASR) is the area of research dealing with the design and development of systems concerned with this problem (Markowitz, 1996). Despite recent incredible progress, ASR systems are still far from being perfect and depending on the conditions and environment in which the speech was uttered, have recognition accuracy measured as "word error rate"(WER) ranging between 10 and 60 percent. Obviously the effectiveness of spoken queries is tied to their accurate recognition by ASR (Crestani, 2000; Barnett, Anderson, Broglio, Singh, Hudson, & Kuo, 1997).

The advantages of speech as a way to express an information need are obvious. It is natural just as people communicate as they normally do; it is fast: commonly 150-250 word per minute (Aronson & Colet, 1997); it requires no visual attention; it requires no use of hands. All mobile phones and many PDAs are equipped with microphones and that could become IR terminals. However, ASR is imperfect, which means that there is bound to be recognition mistakes at different levels depending on the quality of the ASR systems. Queries are generally short. The shorter duration of spoken queries provides less context and redundancy for ASR, and errors will have a greater impact on effectiveness of IR systems (Allan, 2001; Crestani, 2000). In addition, spoken queries need to be processed online and "almost" in real time. This intensifies the already computational expensive recognition process and demands the time for speech

process to be kept short. 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 imposed by speaking must not be ignored. Generally when formulating spoken queries, users are not simply transcribing information but are composing it. For such tasks, the real limiting factor may be how quickly one can generate and formulate ideas. In this sense, it is no different from an accomplished typist who may be able to copy information quickly, but is slowed down considerably when having to compose original text. However, despite the unavoidable ASR errors, voice is more expressive. People express themselves more naturally and less formally when speaking compared to writing and are generally more personal. It has long been proved that voice is a richer media than written text (Chalfonte, Fish, & Kraut, 1991). Thus, we would expect that given the user information need, in general a spoken query would be longer in length than a written query. Furthermore, the translation of thoughts to speech is faster than the transition of thoughts to writing. So the process of formulating a spoken query should be shorter than that of formulating a written query. To test these two ideas, we carried out an experiment as described in the section 2. However, before describing our work, we are going to put it in the context of related work.

2.2 Related Work

During the last two decades, very-large-vocabulary speech recognition tech-

nology has been successfully developed and has been incorporated into some IR systems.

Spoken query processing (SQP), sometime also referred to as speech-driven information access, is concerned with retrieving documents in response to a spoken query. The emphasis is on the query that is spoken, documents can either be written or spoken. This area of research should not be confused with spoken document retrieval (SDR), which is instead concerned with retrieving spoken documents in response to a written query. From 1997 (TREC-6) to 2000 (TREC-9), the TREC (Text REtrieval Conference) evaluation workshop included a track on SDR to explore the impact of ASR errors on document retrieval. The conclusion draw from these three years of SDR track is that SDR is almost a ”solved problem” (Garofolo, Auzanne, & Voorhees, 1999).

Conversely, very little work has been devoted to SQP. So far, SQP has very much 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. Kupiec et al. used a speaker-dependent speech-recognition system to recognise spoken keywords for information retrieval. A hypothesised phone sequence was generated from the speech-recognition system where each input keyword is spoken in isolation. The hypothesised phone sequence was also matched against possible keywords in the document database. Kupiec reported satisfactory results when the system was used to query articles in an encyclopaedia (Kupiec, Kimber, & Balasubramanian, 1994). Research carried out by Barnett et al. showed that longer queries are more robust in terms of

tolerating errors than shorter queries, although increasing WRE does result in decreasing precision (Barnett et al., 1997). Crestani carried out a study to investigate further on the effects of WRE and query length (Crestani, 2002). The study underpinned the previous conclusion by Barnett et al. Moreover, it also concluded that both standard relevance feedback and pseudo relevance feedback enable to improve the effectiveness of SQP, in particular for short queries. Fujii and his colleagues showed that using a language model generated from the target collection can significantly improve both the recognition and retrieval accuracy (Fujii, Itou, & Ishikawa, 2002). However, these studies focused solely on investigating the effects of speech recognition accuracy on IR methods based on non-spontaneous (i.e. read) and long queries and did not take into account the major properties of IR during the searching process, such as the effects of different query interfaces on the performance of IR systems. In fact, SQP is more complicated. It involves the integration of an ASR system and an IR system, and is not ”simply connected by way of an input/output protocol” to an IR system (Fujii et al., 2002). This view was taken by (Chien, Wang, Bai, & Li, 2000) who built an efficient spoken-access approach for both Chinese text and Mandarin speech information retrieval, enabling users to submit spoken queries in an interactive way. The extensive experimentation reported in this paper shows that speech interaction can improve the effectiveness of information retrieval. However, while these conclusions support our line of research, it is not clear what the retrieval effectiveness of spoken queries is when no interaction or performance enhancement techniques (e.g. relevance feedback) are involved.

We consider this to be the baseline of the effectiveness of spoken queries and we feel it should be assessed and compared with the baseline given by written queries. This is what we set up to investigate.

3 Comparison of Written versus Spoken Queries

This section presents the experimental procedure, the results and the analysis of a study on the qualitative and quantitative differences between spoken and written queries. The latter is an extended version of the results and analysis reported in (Du & Crestani, 2003).

Table 1: Characteristics of written and spoken queries.

3.1 Experimental Procedure

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 users in a setting as close as possible to a real world application. We used a within-subjects experimental design (Miller, 1984), including 12 users. We are aware that this is a small user sample, but this is still, to our knowledge, the largest study of its kind. In addition, in accordance with Nielsen et al., we think that if the experimental procedure is well designed we do not need a large

user sample (Nielsen & Landauer, 1993)

As retrieving information via voice is still relatively in its infancy and is not a well known technology, we recruit potential users from an accessible group who was not new to the subject of IR. Seven of our participants were from the local IR research group who have knowledge of IR to some degree and five participants were research students who all have good experience of using search engines within our department. Our subjects participated in the experiment voluntarily and were all native English speakers.

The topics we used for this experimental study from which queries were constructed were a subset of 10 topics extracted from TREC topic collection. Each topic consists of four parts: id, title, description and narrative. An example of such topic is shown here, slightly reformatted from the original version for legibility:

<title> Topic: Coping with overcrowded prisons

<desc> Description: The document will provide information on jail and prison overcrowding and how inmates are forced to cope with those conditions; or it will reveal plans to relieve the overcrowded condition.

<narr> Narrative: A relevant document will describe scenes of overcrowding that have become all too common in jails and prisons around the country. The document will identify how inmates are forced to cope with those overcrowded conditions, and/or what the Correctional System is

doing, or planning to do, to alleviate the crowded condition.

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 descriptions of 10 TREC topics in text form. The 10 topics were in a predetermined order and each had a unique ID. The tasks were that each participant was asked to formulate a query for each topic in either written form or spoken form as instructed via a graphic user interface (GUI) on a desktop screen. For session 1, each participant was asked to formulate queries in written form for the first 5 queries and in spoken form for the second 5 queries. For session 2, the order was reversed, that is each participant formulated queries in spoken form for the first 5 topics and in written form for the second 5 topics. A maximum of 5 minutes time constraint was imposed on each topic. Session 2 was carried out one week after session 1, this was because after the participants had taken part in session 1, they had familiarised themselves with the 10 topics to some degree, which would pose a threat to the validity of our data if they worked with the same topics in session 2 immediately. By running session 2 some time after session 1, we hoped this threat would be minimised. At the end of the experiment, each participant was interviewed for about 10 minutes and a questionnaire was presented to each participant in order to obtain additional

information about the query formulation process.

We utilised three different methods of collecting data for post-experimental analysis: background system logging, interviews and questionnaires. Through these means we could collect data that would allow us to analyse and test the experimental hypotheses. During the course of the experiment, the written queries were collected and saved in text format along with the duration of the formulation process. The duration of each written query was considered as the total time a participant spent to comprehend a topic, formulate the query in the query field and submit it using the submit button in the GUI. Spoken queries were recorded and saved in audio format in a wav file for each participant automatically along with the duration for each query. After reading a topic, to record a query, the participant could click the ”start speaking” button and speak the query into a microphone and then click ”stop speaking” to terminate the recording. So, the duration of each spoken query was calculated as the total time a participant needed to comprehend a topic and record the query. The interviews sought to solicit participants’ comments on the GUI design and explanations for the occurrence of some exceptional behaviour the experimenter observed during the course of experiment. Participants were also asked to point out the easiest and most difficult topics in written and spoken form and the reasons for their judgments. The same questionnaires would be handed out after the completion of both sessions to gather participants’ assessment on the complexity of the tasks. By comparing their answers, we could see how their ratings on the difficulty of the tasks would vary from session 1 to session 2.

We also wanted to evaluate the difference in retrieval effectiveness between written and spoken queries. For this a suitable test environment needed to be devised. Classical IR evaluation methodology (van Rijsbergen, 1976) suggests that such test environment should consist of the following components: (a) a collection of textual document; (b) a set of queries with associated document relevance assessments; (c) one or more IR systems; and (d) some measures of IR system e□ectiveness.

The collection we used is a subset of the collection generated for TREC (Voorhees & Harman, 1998). The collection consists of the full text of articles of the Wall Street Journal from year 1990 to year 1992 and comprises about 74.000 documents. The 120 written and 120 spoken queries collected from above mentioned experiment were used to retrieve document from the collection. 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 IR 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 (Ogilvie & Callan, 2001). It supports indexing of large text collection, the construction of simple language models for documents and queries and the implementation of IR systems based on a variety of retrieval models.

The 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, precision and recall can be easily evaluated. The results were averaged over the entire sets of 120 written queries and 120 spoken queries.

3.2 Results and analysis

In this section we present the results and analysis of the heuristic evaluation carried out with the procedure previously described.

3.2.1 Query Length

Using the described procedure, we collected 120 written queries and 120 spoken queries. Some of the characteristics of written and spoken queries are reported in Table 1. These data show clearly that the average length of spoken queries is longer than written queries as we have hypothesised with a ratio rounded at 2.4.

The average length of spoken and written queries for each topic across all 12 participants is presented in figure 1. Notice how the markers for spoken queries are always above the markers for written queries, which indicates that spoken queries were almost always lengthier than the written ones for any topic. This was exactly what we expected to see. We know from previous studies that textual queries posed to information retrieval systems by untrained users are short: most queries are three words or less. With some knowledge of information retrieval and high usage of web search engines, our participants formulated longer textual queries. When formulating queries verbally, the ease of speech encouraged participates to formulate longer queries. 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’m 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 is much shorter: ”Right wing Christian fundamentalism, grass roots, civil liberties, US Constitution.”

Figure 1: Average length of queries per topic.

We also summarised the length of queries for all 10 topics across all participants. The average length of queries per participant is presented in figure 2. We can observe from figure 2 that spoken queries were longer than written ones consistently for every participant. However, the variations of the length between spoken and written queries for some participants were 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 were almost identical to their written ones. In this case, the discrepancies of length within written queries were very insignificant and relatively stable since all participants used similar approach to formulate their written queries by specifying only keywords. Conversely, the length fluctuated widely within spoken queries among participants. In this experiment, we observed that 8 out of 12 participants adopted proper natural language to formulate their spoken queries which were very much like

conversational talk, while 4 participants only spoke keywords and/or broken phrases. They commented that they didn't want to "talk" to the computer, because they felt strange and uncomfortable to speak to a machine. They just "spoke written queries". This was due to their knowledge of how most IR systems and search engines work. Since they knew that these systems often remove stopwords, they did not see any point in formulating a spoken natural language statement and used the same keyword-based approach to formulate spoken queries that they used to formulate written queries.

Figure 2: Average length of queries per participant.

3.2.2 Query Duration

The time spent to formulate each query was measured. A maximum of 5 minutes was imposed on each topic and all participants felt that the time given was sufficient. There was only one occasion a participant didn't formulate a written query within the time limit and this was due to a topic that appeared particularly ambiguous to the participant.

Figure 3: Average duration of queries per topic.

The average time participants spent on each topic is shown in figure 3. For the first half topics, more time was needed to form the written queries than spoken ones but the discrepancy was not as great as we expected. Participants

spent almost the same time to formulate query in written and spoken forms for each of the second half topics. From this figure, we were able to establish that no significant difference existed between the two query forms in terms of the duration. This appears to counter our claim that participants would require less time to form spoken queries than written queries. However, we cannot neglect the fact that the cognitive load on participant to speak out their thoughts was also high. Some of them commented that they had to formulate well their queries in their head before speaking aloud with no mistakes. Speech driven information access is a relatively new research area and there aren't many working systems available currently. We believe lack of experience put more pressure on the spoken query formulation process.

Figure 4: Average duration of queries per participant.

The duration of queries per participant is shown in figure 4. Some participants spent less time on spoken queries than written ones, whereas it was a reverse case for some other participants. The variations of durations across all participants were very irregular and there were no significant differences among the durations for the two forms, therefore, we were unable to establish any strong general claim. Nevertheless, the figure did show that two thirds of the participants spent less time on spoken queries than written ones whereas only one third of the participants required more time for spoken queries than written ones.

3.2.3 Query Length without stopwords

From the previous analysis, we know that spoken queries in general are lengthier than written queries. One would argue that people with natural tendency to speak more conversationally would formulate spoken queries as long sentences containing a great deal of function words such as prepositions, conjunctions or articles that have little semantic contents of their own. Such words have been referred as stopwords in IR and often discarded from document and query representations. So we removed the stopwords from both spoken and written queries and plotted the average length of spoken and written queries against their original length.

Figure 1 show query length after stopwords removal. The average length of spoken queries reduced from 23.07 to 14.33 with a 38% reduction rate, while the average length of written queries reduced from 9.54 to 7.48 with a reduction rate at 22%. These figures indicated that spoken queries contained more stopwords than written ones. This indication can also be seen from difference between the average length and median length for both spoken and written queries.

Figure 5: Average length of queries across topics.

As we can see from figure 5, the markers for spoken queries is consistently on top of the ones for the written queries for every participants, even after stopword removal, though spoken queries are undoubtedly becoming shorter. Moreover, the markers for spoken queries without stopwords stay above the ones for written queries without stopwords consistently also across every topic as

depicted in figure 6. Statistically, the average spoken query is still almost double the length of the written ones. This significant difference in length indicates that the ease of speaking encourages people to express not only more conversationally, but also more semantically.

Figure 6: Average length of queries per participant.

3.2.4 Part of speech

A natural language sentence is usually composed of nouns, pronouns, articles, verbs, adjectives, adverbs, connectives, etc. From the 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. Figure 7 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 different 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. Nouns and nouns phrases are well known to carry more information content and therefore more useful for search purposes. The fact that they are less frequent in spoken queries could be detrimental to their

effectiveness. 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.

Figure 7: Percentages of part-of-speech in written and spoken queries.

3.2.5 Retrieval Effectiveness

After having discovered substantial qualitative differences between written and spoken queries, it is now worth studying if these differences make any impact on the effectiveness of written and spoken queries in the information retrieval task. In other words, given the same information need, would a written query be generally more effective in identifying relevant documents than a spoken one?

This section describes 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 ASR process is perfect. This is of course a gross simplification, since even well trained ASR systems make recognition errors. Nevertheless, this study could provide the upper bound level of performance of an IR system using spoken queries.

We ran the written and spoken sets of queries against three retrieval models implemented using the Lemur toolkit: a basic tf-idf vector space model, the

Okapi, and a language modelling method that used the Kullback-Leibler similarity measure between document and query language models (Lin, 1991). No relevance feedback methods were used for any of these three models. We used 3 different models to avoid any possible bias a model could have in favour of a specific type of query formulation. We believe the use of these 3 very different models enables us to generalise the results obtained with more certainty.

Notice also that we removed stopwords from both sets of queries, since this is common practice in most IR systems and since the inclusion of stopwords would not change significantly the results of the analysis. Obviously, the same standard list of stopwords was used for both query sets and for all 3 models.

Table 2 depicts the effectiveness of written and spoken queries using the three models. Naturally, we would expect the best result to come from the spoken queries, since they are longer, but the performances obtained for the two query sets are very similar and a paired t-test showed that the difference

Table 2: Precision at 11 standard points of recall for spoken and written queries using all terms in the queries.

is not statistically significant. As the result, we cannot conclude that one is better than the other based on a small performance difference between two query sets.

Figure 8: P/R graph for simple tf-idf 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, that is, each query in this new set comprised of the terms appearing in both the written and spoken corresponding queries. The results obtained with this set of queries, shown in figure 8 for the tf-idf model, indicate that this query set obtained slightly better retrieval performance. This is an indication the important words (those responsible for the retrieval effectiveness) are those that are present in both written and spoken query sets. Those words that are present only in spoken or written queries are therefore responsible for the decrease in performance. The same results were obtained using the other two models.

So, from table 2, we can conclude that these two sets of queries are almost equally effective with respect to retrieval performances, but from our data on written and spoken query length, we could claim that spoken queries are more useful than written queries because they carry more content words. In fact, as far as IR performance is concerned, more content words should lead to more effective relevant document retrieval. So, where have the content words gone during the retrieval process? Our work 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.*

A number of words appearing in the written query are also present in the spoken query. Other words in the 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. The vocabulary sizes of these three query sets are shows in Table 3; 71% of words in written queries are common words whereas only 40.9% of spoken query words are common. The ratio of common terms over the total vocabulary sizes of

written and spoken queries is 25.9%.

Table 3: Vocabulary size of written and spoken queries.

Table 4, 5 and 6 report the retrieval effectiveness of written and spoken queries when only specific part of speech words are used. These tables show that there are no significant differences between the retrieval effectiveness obtained by written and spoken queries when only nouns (table 4), nouns and adjectives (table 5) or nouns, adjectives and verbs (table 6) are used. So, despite substantial differences in percentage of part of speech in written and spoken queries, written queries give the same level of retrieval effectiveness as spoken queries. While this might seem counterintuitive, previous analysis of the overlap between written and spoken queries might still provide an explanation for this behaviour. Terms that are crucial to retrieval effectiveness are present in both written and spoken queries. Terms that are present only in one query set are not so important to retrieval effectiveness. These tables also show that while nouns are indeed important to retrieval effectiveness, the inclusion in queries of adjectives and verbs increases the retrieval effectiveness. Conversely, the inclusion of other parts of speech in both query types has no effect whatsoever on retrieval effeteness. Figure 7 shows that these other parts of speech make up a higher percentage of spoken query terms than written query terms.

Table 4: Precision at 11 standard points of recall for spoken and written queries using only nouns in the queries.

3.2.6 User Experience in Formulating Spoken Queries

An analysis of the post experimental questionnaire filled by participants showed that some of them found very natural to formulate spoken queries, while others found it awkward. The first set of participants corresponds almost exactly with those that expressed their queries as a natural language statement and expressed themselves in a very colloquial way. The second set of participants comprised almost exclusively of those that expressed they queries as ”spoken keywords”. These participants indicated that having to select and keep in mind keywords before speaking them into the system was a very complex task and that they felt more comfortable doing it in writing. They argued that doing this task in writing gave them a chance to add/remove keywords from the query before hitting the submit button in a much easier way than doing it in their mind. This second set of participants was composed mainly by senior male participants, with more experience and knowledge of IR technology than the first set of participants.

Table 5: Precision at 11 standard points of recall for spoken and written queries using only nouns and adjectives in the queries.

Table 6: Precision at 11 standard points of recall for spoken and written queries using nouns, adjectives and verbs in the queries.

Another recurrent comment found in the questionnaires was related to the difficulty of generating queries from TREC topics. This comment was made for both spoken and written queries. Participants also complained that sometime they had almost no knowledge of the topic and had to rely completely and solely on the text of the topic to formulate the query. We obviously expected a comment on this, having used this experimental procedure in other experiments (see for example (Tombros & Crestani, 2000)). However, we had no choice in this matter given the cost of building a test collection with relevance assessments.

4 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 in retrieval performance, assuming perfect transcription of the spoken queries. This study serves as the basis for the design of a speech user interface that will enable access information via spoken queries and spoken interactions. The results of the work reported here 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. However, despite being longer in terms of number of words, spoken queries do not seem to be significantly more effective than written ones, even assuming perfect ASR.

IR systems are very sensitive to query formulation. Errors in queries, produced by imperfect ASR, are a new problem for IR that was never en-

countered before (Allan, 2001). So, while spoken queries and spoken interactions can open very interesting research directions in IR, they also bring new challenges. In the future, we intend to study different ways to take advantage of the distinctive characteristics of spoken medium. In fact, in the study represented here we did not take advantage of any of the important features of speech, like its highly interactive nature, its richness, and its expressiveness. We considered only the characteristics of transcribed speech. However, they are many ways in which these characteristics of speech can be used for more effective IR, like for example:

- A vocal dialogue manager can make interaction between IR system and user more natural and effective, encouraging the use of techniques like relevance feedback and results browsing, as shown in (Chien et al., 2000).
- Language models can be devised that are able to capture in a more effective way the information need expressed in a spoken query.
- Non-verbal information contained in speech, like for example prosodic stress can be used to identity words that are important in characterising the user information need. This has been shown in (Silipo & Crestani, 2000) to improve retrieval performance.

We are currently working in all these directions.

Finally, as a side to this research, we carried 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. Our preliminary analysis of the results, reported in (Du & Crestani, 2004), seems to confirm that the conclusions reached for English and presented here also hold for Mandarin.

References

Allan, J. (2001). Perspectives on information retrieval and speech. *SIGIR Workshop: Information Retrieval Techniques for Speech Applications* (pp. 1–10).

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

Barnett, J., Anderson, S., Broglio, J., Singh, M., Hudson, R., & Kuo, S. (1997). Experiments in spoken queries for document retrieval. *Proceedings of Eurospeech*, Vol. 3 (pp. 1323–1326).

Chalfonte, B., Fish, R. S., & Kraut, R. E. (1991). Expressive richness: a comparison of speech and text as media for revision. *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology* (pp. 21–26).

Chien, L., Wang, H., Bai, B., & Li, S. (2000). A spoken-access approach for Chinese text and speech information retrieval. *Journal of the American Society for Information Science*, *51* (4), 313–323.

Cool, C., Park, S., Belkin, N., Koenemann, J., & Ng, K. (1996). Information

seeking behaviour in new searching environment. *CoLIS*, 403–416.

Crestani, F. (2000, May). Effects of word recognition errors in spoken query processing. *Proceedings of the IEEE ADL 2000 Conference* (pp. 39–47). Washington DC, USA.

Crestani, F. (2002). Spoken query processing for interactive information retrieval. *Data and Knowledge Engineering*, *41* (1), 105–124.

Du, H., & Crestani, F. (2003). Spoken versus written queries for mobile information access. *Proceedings of Mobile HCI 2003 International Workshop on Mobile and Ubiquitous Information Access* (pp. 67–78).

Du, H., & Crestani, F. (2004). Spoken versus written queries for mobile information access: an experiment with mandarin Chinese. *Proceedings of IJCNLP 2004, the 1st International Joint Conference on Natural Language Processing* (pp. 358–365). Sanya City, Hainan Island, China.

Fujii, A., Itou, K., & Ishikawa, T. (2002). Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. *Proceedings of the SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications* (pp. 94–104). Springer LNCS 2273.

Garofolo, J. S., Auzanne, C., & Voorhees, E. M. (1999, November). The TREC spoken document retrieval track: a success story. *Proceedings of the TREC Conference* (pp. 107–130). Gaithersburg, MD, USA.

Kupiec, J., Kimber, D., & Balasubramanian, V. (1994). Speech-based retrieval using semantic co-occurrence filtering. *Proceedings of Human Language*

Technology Conference.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Tran. on Information Theory*, *37* (1), 145–151.

Markowitz, J. (1996). *Using speech recognition*. Prentice Hall, Upper Saddle River, NJ, USA.

Miller, S. (1984). *Experimental design and statistics*. London, UK: Routledge, second edition edition.

Nielsen, J., & Landauer, T. (1993). A mathematical model of the finding of usability problems. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 206–213). Amsterdam, The Netherlands.

Ogilvie, P., & Callan, J. (2001). Experiments using the lemur toolkit. *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)* (pp. 103– 108).

Silfverberg, M., MacKenzie, S., & Korhonen, P. (2000). Predicting text entry speed on mobile phones. *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems* (pp. 9–16). The Hague.

Silipo, R., & Crestani, F. (2000). Prosodic stress and topic detection in spoken sentences. *Proceedings of the SPIRE 2000, the Seventh Symposium on String Processing and Information Retrieval* (pp. 243–252). La Corunna, Spain.

Soukoreff, W., & MacKenzie, I. (1995). Theoretical upper and lower bounds on typing speeds using a stylus and keyboard. *Behaviour and Information Technology*, 379–379.

Tombros, A., & Crestani, F. (2000). User's perception of relevance of spoken

documents. *Journal of the American Society of Information Science*, *51* (9), 929–939.

van Rijsbergen, C. (1976). *Information retrieval*. London, UK: Butterworths, second edition edition.

Voorhees, E., & Harman, D. (1998). Overview of the seventh text retrieval conference (TREC-7). *Proceedings of the TREC Conference* (pp. 1–24). Gaithersburg, MD, USA.

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

Table 1: Characteristics of written and spoken queries.

Recall	Prec. TF-IDF		Prec. Okapi		Prec. KLJM	
%	Spoken	Written	Spoken	Written	Spoken	Written
0.00	0.70	0.65	0.74	0.68	0.72	0.63
0.10	0.57	0.56	0.63	0.61	0.60	0.56
0.20	0.46	0.48	0.55	0.53	0.52	0.49
0.30	0.42	0.43	0.48	0.47	0.46	0.44
0.40	0.34	0.37	0.39	0.39	0.39	0.38
0.50	0.31	0.34	0.34	0.35	0.32	0.32
0.60	0.22	0.25	0.24	0.28	0.22	0.24
0.70	0.17	0.19	0.17	0.21	0.16	0.18
0.80	0.15	0.16	0.14	0.17	0.14	0.15
0.90	0.10	0.11	0.08	0.09	0.09	0.10
1.00	0.06	0.07	0.05	0.07	0.07	0.08

Table 2: Precision at 11 standard points of recall for spoken and written queries using all terms in the queries

	Written queries	Spoken queries	Common query terms
Vocabulary size	309	552	226

Table 3: Vocabulary size of written and spoken queries.

Recall	Prec. TF-IDF		Prec. Okapi		Prec. KLJM	
%	Spoken	Written	Spoken	Written	Spoken	Written
0.00	0.55	0.56	0.60	0.58	0.55	0.47
0.10	0.42	0.44	0.47	0.48	0.41	0.40
0.20	0.34	0.37	0.39	0.39	0.35	0.33
0.30	0.29	0.33	0.32	0.34	0.29	0.29
0.40	0.24	0.29	0.27	0.30	0.25	0.25
0.50	0.21	0.26	0.22	0.25	0.19	0.20
0.60	0.14	0.17	0.14	0.16	0.14	0.14
0.70	0.11	0.11	0.09	0.11	0.10	0.10
0.80	0.09	0.09	0.07	0.08	0.08	0.08
0.90	0.05	0.05	0.03	0.02	0.04	0.04
1.00	0.03	0.04	0.01	0.02	0.03	0.03

Table 4: Precision at 11 standard points of recall for spoken and written queries using only nouns in the queries.

Recall	Prec. TF-IDF		Prec. Okapi		Prec. KLJM	
%	Spoken	Written	Spoken	Written	Spoken	Written
0.00	0.64	0.61	0.70	0.63	0.67	0.57
0.10	0.49	0.51	0.57	0.56	0.53	0.50
0.20	0.41	0.43	0.48	0.48	0.44	0.42
0.30	0.37	0.38	0.42	0.43	0.40	0.39
0.40	0.30	0.33	0.35	0.36	0.34	0.33
0.50	0.27	0.30	0.30	0.32	0.28	0.27
0.60	0.19	0.21	0.20	0.23	0.20	0.19
0.70	0.14	0.14	0.14	0.16	0.14	0.14
0.80	0.12	0.11	0.11	0.13	0.12	0.12
0.90	0.07	0.07	0.05	0.06	0.08	0.07
1.00	0.05	0.05	0.03	0.05	0.06	0.06

Table 5: Precision at 11 standard points of recall for spoken and written queries using only nouns and adjectives in the queries.

Recall	Prec. TF-IDF		Prec. Okapi		Prec. KLJM	
%	Spoken	Written	Spoken	Written	Spoken	Written
0.00	0.69	0.66	0.74	0.68	0.71	0.64
0.10	0.56	0.56	0.63	0.61	0.60	0.57
0.20	0.46	0.48	0.54	0.53	0.51	0.49
0.30	0.41	0.43	0.47	0.48	0.45	0.45
0.40	0.34	0.37	0.38	0.40	0.38	0.39
0.50	0.30	0.34	0.33	0.35	0.32	0.32
0.60	0.22	0.25	0.23	0.28	0.22	0.24
0.70	0.17	0.19	0.18	0.21	0.17	0.18
0.80	0.15	0.16	0.14	0.17	0.15	0.15
0.90	0.10	0.11	0.08	0.09	0.10	0.10
1.00	0.06	0.07	0.05	0.07	0.07	0.08

Table 6: Precision at 11 standard points of recall for spoken and written queries using nouns, adjectives and verbs in the queries.

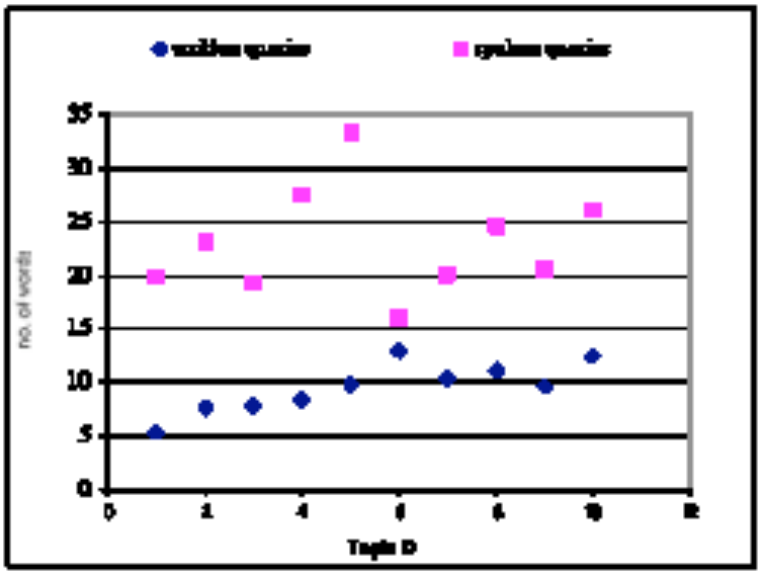

Figure 1: Average length of queries per topic.

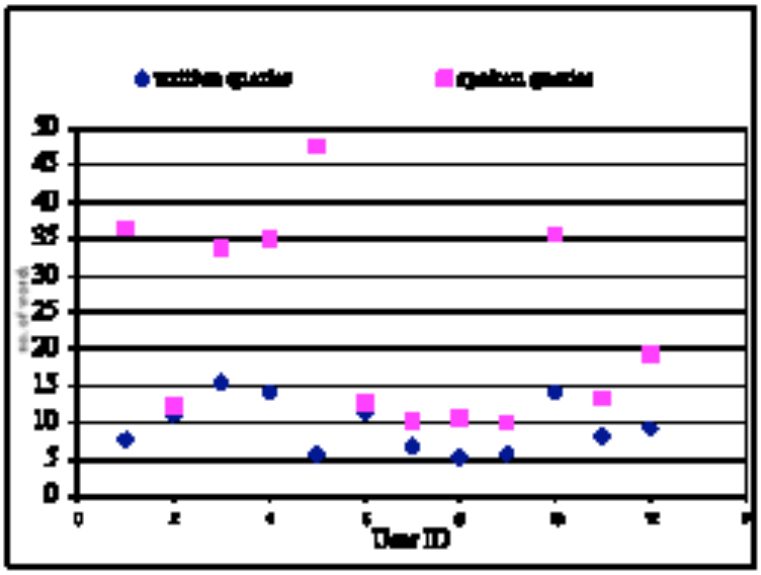

Figure 2: Average length of queries per participant.

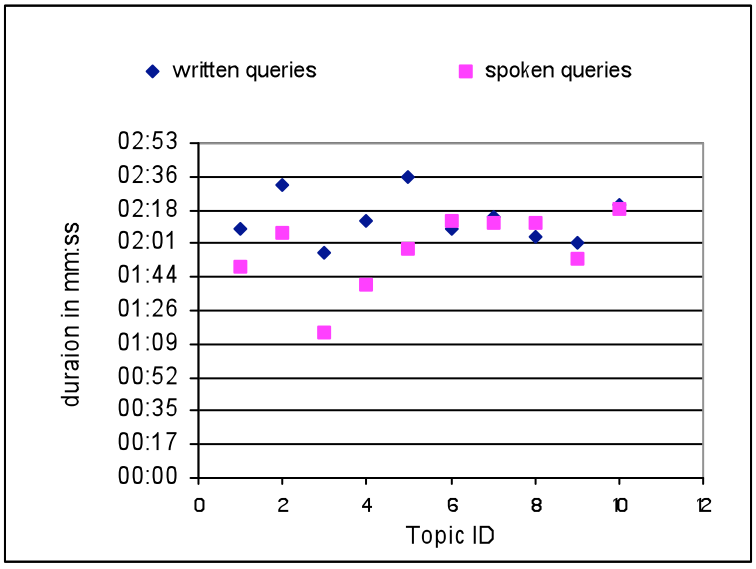

Figure 3: Average duration of queries per topic.

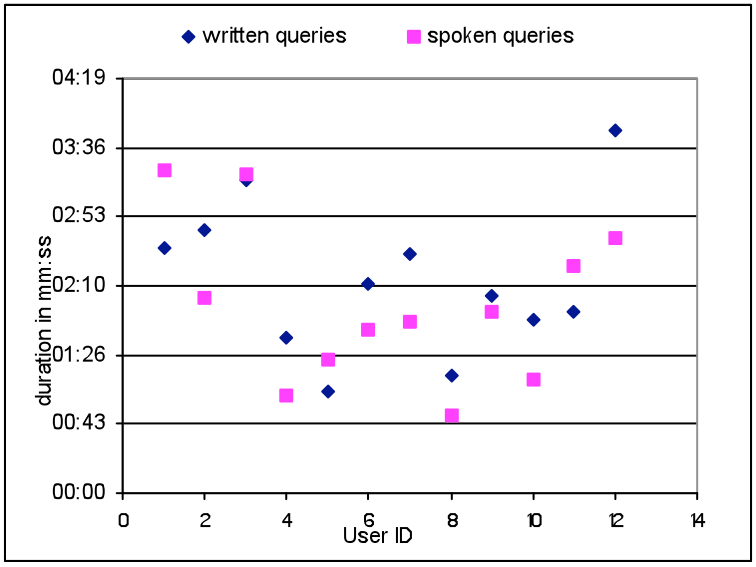

Figure 4: Average duration of queries per participant.

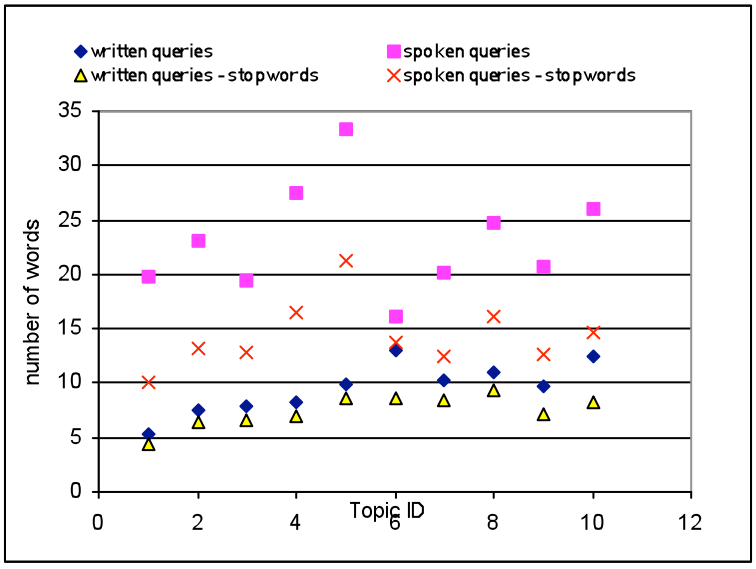

Figure 5: Average length of queries across topics.

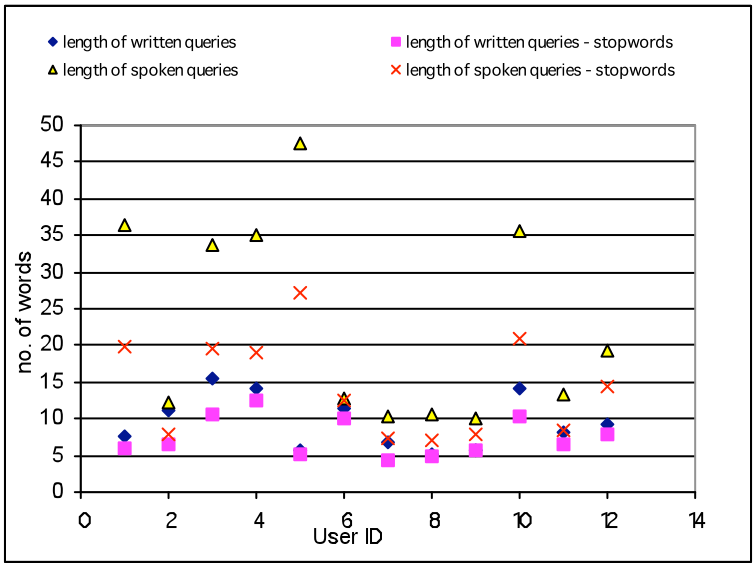

Figure 6: Average length of queries per participant.

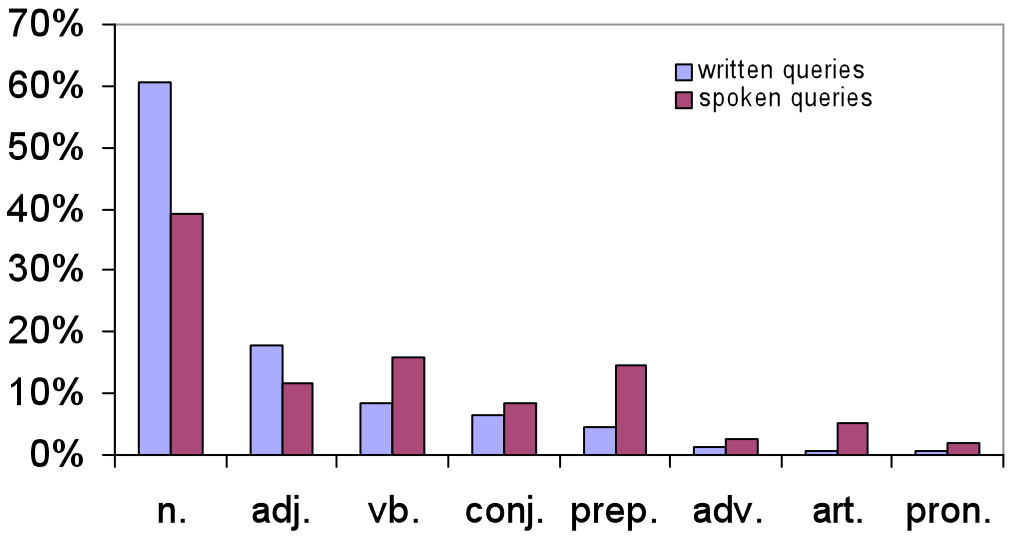

Figure 7: Percentages of part-of-speech in written and spoken queries.

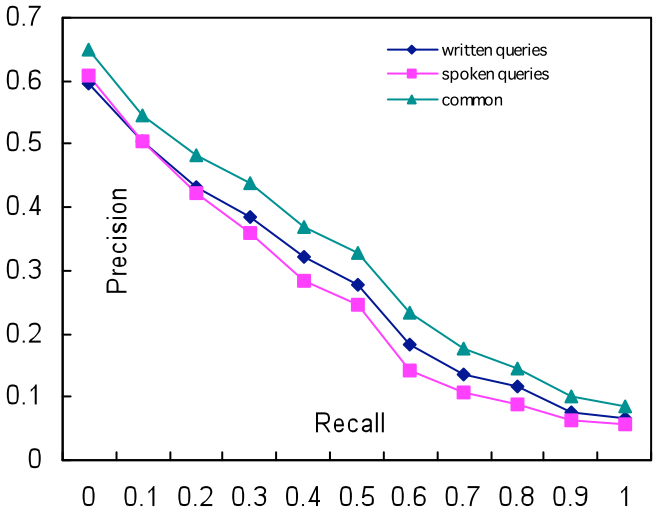

Figure 8: P/R graph for simple tf-idf model