Users' Perception of Relevance of Spoken Documents

Tassos Tombros*
Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland.
E-mail: tombrosa@dcs.gla.ac.uk

Fabio Crestani
International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704.
E-mail: fabioc@icsi.berkeley.edu

We present the results of a study of user's perception of relevance of documents. The aim is to study experimentally how users' perception varies depending on the form that retrieved documents are presented. Documents retrieved in response to a query are presented to users in a variety of ways, from full text to a machine spoken query-biased automatically-generated summary, and the difference in users' perception of relevance is studied. The experimental results suggest that the effectiveness of advanced multimedia Information Retrieval applications may be affected by the low level of users' perception of relevance of retrieved documents.

1. Introduction

There has been a surge of interest in *ubiquitous computing* over the past few years. Ubiquitous computing is an attempt to break away from the traditional desktop interaction paradigm by distributing computational power and resources into the environment surrounding the user. In the last few years, there has also been an increasing emphasis on extending the utility of information systems by providing access through mobile devices, for example, telephones or PDAs (Goose et al., 1998). Enabling access to an information service via a telephone, without the use of a computer and a modem or a dedicated client device, has the potential to considerably increase the size of the user community. In addition to offering greater convenience and flexibility, ubiquitous access to information services via telephone devices enables professionals to make use of previously unproductive time. Moreover, the use of audio input and output enables visually impaired users to access information services without any of the problems encountered using a computer. In a telephone-based information retrieval (IR) system, the main medium of communication would be vocal.

The introduction of speech in the IR process poses a number of challenges. The challenges are of adifferent nature depending on the context in which speech is introduced. We can have the retrieval of spoken documents using textual queries, the retrieval of textual documents using spoken queries, or, finally, the combination of both (with which we are not going to be concerned here). The retrieval of spoken documents using a textual query is a fast emerging area of research (see, for example, Sparck-Jones et al., 1996). It involves an efficient, rather than effective, combination of the most advanced techniques used in speech recognition and IR. The increasing interest in this area of research is confirmed by the inclusion, for the first time, of a "retrieval of spoken documents" track in the TREC-6 conference (Voorhees et al., 1997). The challenge is to devise IR models that can cope with the large number of errors inevitably found in the transcripts of spoken documents. Models designed for retrieval of OCRed documents have proved useful in this context (Mittendorf & Schauble, 1996).

Retrieving textual documents using a spoken query may seem easier, because of the smaller size of the speech recognition task involved. However, it is not so. Although the incorrect or uncertain recognition of an instance of a word in a long-spoken document can be compensated for by its correct recognition in some other instances, the incorrect recognition of a word in a spoken query can have disastrous consequences. Queries are generally very short[1] and failure to recognise a query word, or worse, the incorrect recogni-

*To whom all correspondence should be addressed.

Received August 5, 1999; revised February 16, 2000; accepted February 16, 2000.

[1]There is an ongoing debate about realistic query lengths. Although TREC queries are on average about 40 words long, Web queries are only two words long on average. This recently motivated the creation in TREC of a "short query" track, to experiment with queries of more realistic length.

tion of a query word, will fail to retrieve a large number of relevant documents or wrongly retrieve a large number of nonrelevant documents.

Therefore, enabling access to an IR system via a telephone is a much more complex task than one may think. The low bandwidth offered by a telephone line and the level of noise present in many telephone services create a series of additional problems. First of all, the system may have difficulties in recognising the user's commands and queries. The system will need to be capable of interacting with the user, assisting him to clarify and specify his information need. Moreover, the user may find it difficult to understand the response of the system and may not be able to use it as efficiently as a conventional on-screen IR system. These difficulties need to be addressed to be able to implement such a system effectively.

This paper is concerned with the last of these issues: evaluating the effectiveness of a telephone-based IR service from the user's perspective. In particular, we addressed and studied the effectiveness of the users' perception of the relevance of document summaries presented via a vocal interface. This issue is very important for the correct assessment of the feasibility of a telephone access to an IR system. The particular aspect of relevance we are examining is that of topicality (Schamber et al., 1990). Topicality can be defined as the relation of a document to the topic of a user's query, i.e. "relevance to a subject", in the words of Vickery (Vickery, 1959). According to this view, and in the context of this paper, user's perception of relevance should be interpreted as user's perception of topicality. Although it is widely recognized to be an important component of a relevance decision, user's perception of topicality has been little explored or studied.

The paper is structured as follows. Section 2 describes the background and motivations of this study. Section 3 reports some considerations on previous studies of the user's perception of relevance. The core of the paper follows, starting with a description of the experimental system employed in this study, reported in Section 4. The experimental design of our user study is reported in Section 5, and the results are described and analysed in Section 6. Finally, Section 7 reports the conclusions of our work and points at directions of future extensions of this study.

2. Background

The background of the work reported in this paper is related to a project currently underway at the University of Glasgow: the Sonification of an Information Retrieval Environment (SIRE) project.[2] The main objective of the project is to enable a user to interact with a probabilistic IR system over a low bandwidth communication channel. The next two sections describe the overall goal of the SIRE project and how the study reported in this paper fits into it.

2.1. The SIRE Project

The main objective of the SIRE project is to enable a user to interact (i.e., submit queries, commands and relevance assessments, and receive summaries of retrieved documents) with a probabilistic IR system over a low bandwidth communication line (e.g., a telephone line). An outline of the system specification of the prototype is reported in Figure 1.

The prototype interactive vocal information retrieval system (IVIRS) is made up of the following components:

- a vocal dialog manager (VDM) that provides an "intelligent" speech interface between user and IR system;
- a probabilistic IR system (PIRS) that deals with the probabilistic ranking and retrieval of documents in a large textual information repository;
- a document summarisation system (DSS) that produces summaries of the content of retrieved documents in such a way that the user will be able to assess their relevance to his information need;
- a document delivery system (DDS) that delivers documents on request by the user via electronic mail, ftp, fax, or postal service.

It is important to emphasise that such a system cannot be developed simply with off-the-shelf components. In fact, although some components (DSS, DDS, and the Text-to-Speech module of the VDM) have already been developed in other application contexts, it is necessary to modify and integrate them for the IR task.

The IVIRS prototype works in the following way. A user connects to the system using a telephone. After the system has identified the user by means of a username and a password (in the present phase we devised a login procedure based on keying in an identification number using a touch-tone), the user submits a spoken query to the system. The VDM interacts with the user to identify the exact part of spoken dialogue that constitutes the query. The query is then translated into text and fed to the PIRS. Additional information regarding the confidence of the speech recognisers is also fed to the PIRS. This information is necessary to limit the effects of wrongly recognised words in the query. An effective interaction between the system and the user can also help to solve this problem. The system could ask the user for confirmation in the case of an uncertain recognition of a word, asking him to re-utter a word or to select one of the possible recognised alternatives.

The PIRS searches the textual archive and produces a ranked list of documents, and a threshold can be used to find the set of documents regarded as likely to be relevant (this feature can be set in the most appropriate way by the user). The user is informed of the number of documents found to be relevant and can submit a new query or ask to inspect the documents found. Documents in the ranked list are passed

[2]The project is funded by the European Commission under the Training and Mobility of Researchers (TMR) scheme of the European Commission Fourth Framework of projects.

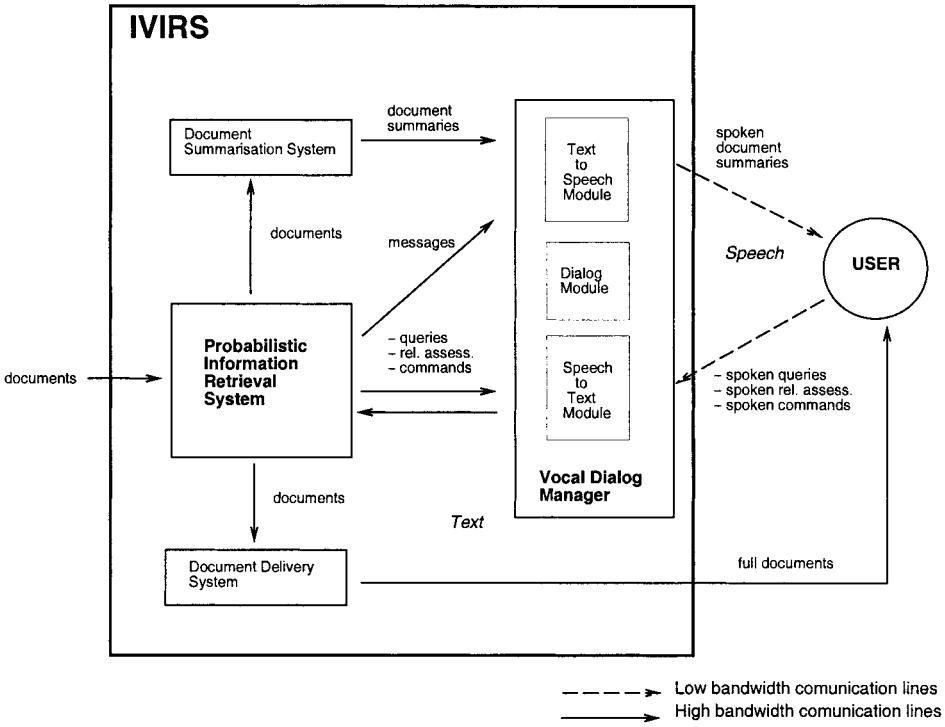

FIG. 1. Schematic view of the IVIRS prototype.

to the DSS that produces a short representation of each document that is read to the user over the telephone by the Text-to-Speech module of the VDM. The user can wait until a new document is read, ask to skip the document, mark it as relevant or stop the process completely.

Marked documents are stored in a retrieved set and the user can proceed with a new query if he wishes to. A document marked as relevant can also be used to refine the initial query and find additional relevant documents by feeding it back to the PIRS. This relevance feedback process is also useful in the case of wrongly recognised query words, because the confidence values of query words may be increased if they are found in relevant documents. This interactive process can go on until the user is satisfied with the retrieved set of documents. Finally, the user can ask for the documents in the retrieved set to be read in their entirety or sent to him via the DDS. The implementation of the prototype system outlined above requires, as a first step, a careful choice of some existing software components: a speech recognition system, a speech synthesis system, a probabilistic IR system, and a document summarisation system. This called for a survey of the state-of-the-art of several different areas of research, some of which are familiar to us, whereas others are new to us. Some components were found not to be fully suitable to the task and had to be developed. This was the case for the probabilistic IR system and the document summarisation system. A second step involves the integration of the various components and the development of a model for the VDM and of its interaction with the other components. Finally, the prototype implementation of the overall system requires a careful tuning and testing with different users and in several different conditions (noisy environment, foreign speaker, etc.).

The prototype implementation of IVIRS is still in progress (Crestani, 1999). A "divide and conquer" approach has been followed, consisting of dividing the implementation and experimentation of IVIRS in the parallel implementation and experimentation of its constituent components. Currently, we have implemented and experimented with the DSS, the Text-to-Speech and Speech-to-Text modules of the VDM, and the DDS. We are currently developing the PIRS (Sanderson & Crestani, 1998), and the VDM (Crestani, 1998b).

2.2. Effectiveness of Spoken Document Summaries

One of the underlying assumptions of the design and development of IVIRS is that a user should be able to assess the relevance of a retrieved document by listening to a synthesised voice reading a brief summary of its semantic content through a noisy channel (e.g., a telephone line). This is obviously essential for an effective use of the system. Moreover, the identification of relevant documents could trigger a relevance feedback process that would not be efficient if fed with nonrelevant documents.

However, results of investigations in other application areas (see, for example, Bernsen et al., 1997; Peckham, 1991) showed that this assumption is not always valid. We therefore decided to carry out a user study aimed at analysing the user's perception of relevance of retrieved documents when these are presented in different forms, and with

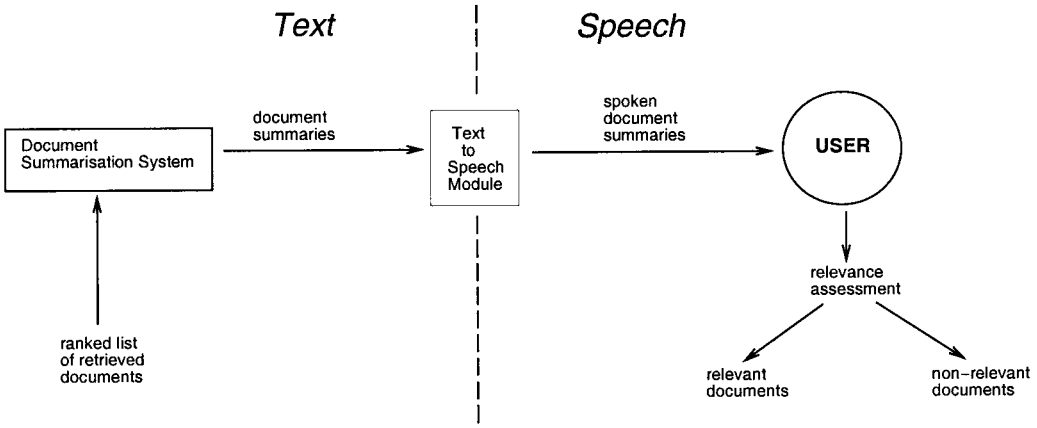

FIG. 2. Generation of spoken document summaries of retrieved documents.

varying levels of distracting elements and noise. The purpose of this study is to evaluate the effectiveness of the delivery of spoken document summaries to the user, an important part of the IVIRS prototype system depicted in Figure 2. In this study, documents retrieved in response to a query will be summarised by our DSS and the summary will be delivered to the user in various forms via different types of Text-to-Speech modules. We aim at evaluating the ability of the user to assess the relevance of the documents whose summaries are being read to him.

3. Users' Perception of Relevance

A user, with an information need expressed in the form of a query submitted to an IR system, may find some information stored in some documents of a document collection "relevant" to his need. In other words, information contained in relevant documents might help the user progress toward satisfying his information need. The goal of an IR system is to retrieve, in response to a query, all and only the relevant documents. To do so, an IR system should be able to identify what makes a document relevant to an information need. It is the ability to capture the characteristics of relevance that enables an IR system to make the difficult decision about what to retrieve and what not to retrieve in response to a query. Thus relevance is one of the most fundamental, if not "the fundamental," concept encountered in the theory of IR, and the notion of relevance, whatever that may be, lies at the heart of the IR process.

In spite of the fact that the concept of relevance is central to IR, and in spite of numerous research attempts to precisely define it, a single satisfactory definition has not yet been given (Mizzaro, 1997). Currently, there are two main views of relevance in IR:

- topic-appropriateness, or topicality, which is concerned with whether or not a piece of information is on a subject which has some topical bearing on the information need expressed by the user in the query;
- user-utility, which deals with the ultimate usefulness of the piece of information to the user who submitted the query.

In current IR research, the term relevance seems to be used loosely in both senses, in spite of the fact that the above distinction is widely accepted. In this paper, we are mainly concerned with the first notion of relevance, namely topicality. This notion is only part of the concept of relevance, but it is the central part in terms of IR evaluation due to its practicality, operational applicability and measurability (Schamber et al., 1990). Research into the concept of relevance has indicated that topicality plays a significant role in the determination of relevance (Saracevic, 1970), although topicality does not automatically result in relevance for users (Barry, 1995). In the same study Barry indicated that motivated users evaluating the relevance of documents would base their evaluation on factors beyond the topical appropriateness of documents. In our experiments, given the fact that we could not use motivated users due to the complexity and scale of the study, we had to resort to topical appropriateness as perceived by users. Taking this contentious view, in this paper we are interested in evaluating how the user's perception of document topicality is affected by the way that the semantic content of the document is presented.

Cuadra and Katter have shown that human relevance judgements are affected by a number of variables (Cuadra & Katter, 1967) that could be grouped into six classes: people, documents, statements of information requirements, judgement conditions, form of response, and judgmental attitudes. Here we are concerned with the judgement conditions and the form of response. Judgement conditions refer to all the external conditions that could affect a user's perception of relevance of a document. These are, for example, the time available for judging a document, or the order in which documents are presented. Form of response refers, according to the original definition given by Cuadra and Katter, to the form in which retrieved documents are presented to the user, for example, title and abstract, full text, or a short summary. Extending this definition to a multimedia and

multimodal IR environment, we could also include different ways of presenting the documents, for example, audio or text. The research reported in this paper investigates the accuracy and speed of user judgements of document topicality when the interaction with the IR system is mediated by an auditory interface, and when documents are presented by means of short, automatically-produced, query-biased summaries.

4. Evaluation of the User's Perception of Relevance of Documents Using Automatically-Generated Spoken Summaries

In this section, we present in detail the two major components of the experimental system depicted in Figure 2: the Document Summarisation System and the Text-to-Speech module.

4.1. Query-Oriented Document Summarisation

Enabling access to an IR service via a telephone, using a vocal interface poses a series of problems. One of these issues is the cost of accessing such a service, and the time needed to interact with the system using vocal commands and responses. It is known that users can rapidly assess the relevance of retrieved documents if they are reading (or skimming) the full text of the articles (Arons, 1997). In a telephone-based IR service, where the communication between the user and the system is performed via a vocal interface, it would be time-consuming and costly to read the full text of the retrieved documents to the user. Moreover, even if the user was not concerned with time and cost, the very nature of the documents may have a confusing effect on the user's ability to assess their relevance: documents may be long and relevant information may be widely scattered, and therefore hard for the user to extract.

It therefore becomes necessary to use shorter versions of the retrieved documents; short enough to be efficiently read over the phone, but indicative enough to enable the user to assess both accurately and quickly the relevance of the documents. It is our belief that the above two requirements could be sufficiently met through the application of query-biased document summarisation methods. A document summary conventionally refers to a condensed version of a document that succinctly presents the main points of the original document (Maizell et al., 1971). Query-biased summarisation methods generate summaries in the context of an information need expressed as a query by a user. Such methods aim to identify and present to the user individual parts of the text that are more focused toward this particular information need than a generic, nonquery-sensitive summary. In this way, summaries can serve an indicative function, providing a preview format to support relevance assessments on the full text of documents (Rush et al., 1971).

Query-biased text summarisation is an emerging area of research that had not been addressed until recently. Tombros and Sanderson looked into the application of such methods in information retrieval, evaluating the indicative function of the summaries (Tombros & Sanderson, 1998). Their study showed that users were better able to identify relevant documents when using the summaries than when using the first few sentences of a document. Recently, the TIPSTER funded SUMMAC project (Mani et al., 1998) provided a framework for the evaluation of different types of summarisation systems. As part of that project, a number of query-biased summarisation systems were evaluated by measuring their ability to help users identify documents relevant to a query.

The summarisation system employed in the experiments described in this paper has been developed by Tombros and Sanderson. The system is based on a number of sentence extraction methods (Paice, 1990) that utilise information both from the documents of the collection and from the queries used. A detailed description of the system can be found in (Tombros & Sanderson, 1998); here we shall briefly describe the summary generation process.

The document collection to be summarised comprised news articles of the Wall Street Journal taken from the TREC collection (Harman, 1996). Each individual document of the collection was passed through the summarisation system, and as a result a score for each sentence of each document was computed. This score represents the sentence's importance for inclusion in the document's summary. Scores are assigned to sentences by examining the structural organisation of each document, and by utilising within-document term frequency information. Information from the structural organisation of the documents was utilised in three ways. Terms occurring in the title section of a document were assigned a positive weight (title score) to reflect the fact that headlines of news articles tend to reveal the major subject of the article. In addition, a positive ordinal weight was assigned to the first two sentences of each article, capturing the informativeness of the leading text of news articles. Finally, a heading score was assigned to each one of the sentences comprising a within-article section heading, reflecting the fact that such headings provide evidence about the article's division into semantic units. By using the number of occurrences of a term in a document (term frequency—TF), we can establish a list of "significant" terms for that document (i.e., terms whose TF value is greater than a specific threshold). The summarisation system then locates clusters of significant terms within a sentence, and computes a significance factor for each sentence (Luhn, 1958).

In addition to the scores assigned to sentences, information from the queries that were used in the experiments was also employed in order to compute the overall score for each sentence. A query score was thus computed, intended to represent the distribution of query words in a sentence. The rationale for this choice was that, by allowing users to see the context in which the query terms occurred, they could better judge the relevance of a document to the query. The actual measure of significance of a sentence in relation to a query is derived using a query length normalisation process.

The final score for each sentence is calculated by summing the partial scores discussed above. The summary for each document is then generated by selecting the top-scoring sentences, and outputting them in the order in which they appear in the original document. Summary length was defined to be 15% of the document's length, up to a maximum of five sentences. Such a value seems to be in general agreement with suggestions made by (Brandow et al., 1995; Edmundson, 1964).

We present here an example of two query-oriented summaries that were generated by the system in response to the query "Combating Alien Smuggling":

> Law—Legal Beat: Two Atlanta lawyers are convicted of immigration fraud.
>
> A Fort Worth, Texas, Federal court jury found the lawyers, Douglas Smith and Ronald Staples, guilty of seven counts each of immigration fraud in connection with a scheme to sell phony documents to illegal aliens seeking legal residence in the U.S. In a January indictment against several members of the ring, the government alleged, among other things, that the two attorneys accompanied the undocumented aliens to immigration offices and assisted them in filing the documents. The attorneys were allegedly part of a nationwide ring that sold packets of bogus addresses, employment histories and medical exams for $3500 to $6000 each. Investigators uncovered an extensive smuggling operation that brought illegal aliens into the U.S. through the Caribbean and other points.
>
> Politics & Policy: South Africa's Armscor may face charge of smuggling U.S. military technology.
>
> Federal prosecutors are preparing the first criminal charges accusing Armscor, South Africa's state-affiliated weapons maker, of smuggling sensitive U.S. military technology to Pretoria, according to law enforcement officials. In addition to Armscor, these officials said, the U.S. attorney's office in Philadelphia intends to seek a pair of indictments naming a host of individuals and smaller companies in a case involving illegal export of missile parts, gyroscopes, and other military hardware for South Africa. Believed to be among the most sweeping international arms-smuggling and financial fraud inquiries in recent years, investigators in the U.S. and elsewhere are still trying to unravel what they contend is a passel of front companies, 39 bank accounts, and fraudulent profit reports used to create more than $1 billion in fake defense contracts. Investigators and former associates contend that Mr. Guerin kept up his connections with Pretoria and later used his intelligence ties to help cover up alleged smuggling and financial fraud. Investigators have said that Mr. Guerin's network of shell companies was used to shuttle money around the world and smuggle military equipment to South Africa.

The next section describes the system we used for the text to speech conversion of the summaries.

4.2. The Text to Speech Module

Speech is the most natural and efficient means by which individuals transmit and access information. However, the ability of the listener to understand the message conveyed by the speaker is highly dependent, among other things, on the quality of the speech.

Speech synthesis is concerned with producing speech by machines (Keller, 1994). Often, this takes the form of a text-to-speech system, whereby unrestricted text is transformed into speech. Since most online information is represented as ASCII text, the automatic conversion of text to speech provides a means to present many people with online information using personal computers or other common devices such as telephones and televisions. Text-to-speech synthesis has the further advantage of providing textual information to people who are visually impaired or functionally illiterate.

The Text-to-Speech module of IVIRS should use state-of-the-art technology in speech synthesis (Keller, 1994). We carried out a survey and an initial testing of a number of commercially available speech synthesis systems. Following a careful selection, we decided to use a system that would be representative of the kind of speech synthesis quality available currently on the market. For the experiments reported in this paper we used the Monologue 97 system.[3] Monologue 97 uses the PrimoVox Speech Synthesizer from First Byte. Monologue 97 for Windows 95 and Windows NT is Microsoft SAPI compliant, and includes a variety of English male and female speech fonts. It is capable of speaking all ANSI text that is made available to it from any application that runs in Windows 95 or NT 4.0. The system is quite flexible because it is able to adjust to a variety of voice characteristics (e.g., speed, tone, pitch, etc.).

However, given the limits and the quality of state-of-the-art speech synthesis systems, we also decided to introduce in our experiments what we considered an "upper bound" of the performance of the Text-to-Speech module: a human voice. Document summaries will be read by a human in different conditions to simulate degrading levels of the quality of speech. The details of the experimental procedure are reported in the following section.

5. Experimental Design

The variable we wish to examine through experimentation (the dependent variable) is the effectiveness of user relevance judgements based on the presented document descriptions. The measures we used to examine the variable are the accuracy of the judgements and the speed with which these judgements were made. In the remaining portion of this section, we present the experimental design of our investigation. Details of the experimental conditions are first provided, followed by a description of the tasks that the users had to perform. The group of subjects participating in the experiments is then described, and finally the experi-

[3]Information on the Monologue 97 system can be found on the First Byte Web site: http://www.firstbyte.davd.com/

mental scenario through which we obtained the measures of user performance is detailed.

5.1. Experimental Conditions

The aim of the experiments reported in this paper is to investigate the effects of different forms of presentation of document descriptions (i.e., with varying levels of distracting elements and noise) on users' perception of document relevance. In a previous study, Tombros and Sanderson (1998) used document titles, and automatically generated query-biased summaries as document descriptions, and measured user ability to make fast and accurate relevance judgements. In that experiment, the descriptions were displayed to the user on a computer screen. The results from that study are used in the present experiments, and will be compared to results obtained when users are listening to the document descriptions instead of reading them. Three different methods of auditory transmission are employed in our study: document descriptions are read by a human to the subjects (condition V), read by a human to the subjects over the telephone (condition T), and finally read by a text to speech application over the telephone to the subjects (condition C). By manipulating the level of the independent variable of the experiments (form of presentation), we are able to examine the value of the dependent variable (i.e., user ability to make fast and accurate relevance judgements). We shall show that any variation in user performance between the experimental conditions can be attributed only to changes in the independent variable, because the so-called "situational variables" (e.g., background noise, equipment used, experimenter's behaviour) are held constant throughout the experimental procedure. Such variables can introduce bias in the results if they change systematically between experimental conditions (Miller, 1984).

5.2. Task

To be able to use the experimental results reported in Tombros and Sanderson (1998), the same task was introduced in our design: users were presented with a retrieved document list in response to a query, and had to identify relevant documents for that particular query within 5 min. The information presented for each document was its title and its automatically-generated, query-biased summary. We also used the same set of queries (50 randomly chosen TREC queries), the same set of retrieved documents for each query (the 50 top-ranked documents were presented to each user), and the same document descriptions (titles, and the query-biased summaries) as in Tombros and Sanderson (1998). The documents are a subset of the Wall Street Journal collection of TREC. To get a measure of user performance in relevance judgements, the TREC relevance assessments were used as the standard against which the subjective judgements of the users participating in the experiment were compared (see Section 3 for a discussion on our view of the concept of relevance). In this way, we were able to produce standard recall and precision figures. One should keep in mind that the focus of the study was not to examine the absolute values of precision and recall (i.e., how much our subjects' and the TREC judges' view of document relevance overlaps), but rather to examine the variation of these measures in relation to the different experimental conditions.

Queries were randomly allocated to subjects by means of a draw, but because each subject was presented with a total of 15 queries (five queries for each condition) we ensured that no query was assigned to a specific user more than once.

5.3. Groups of Subjects

A group consisting of ten users was employed. The population was drawn from postgraduate students doing a conversion course in information technology. Their academic background was from various disciplines (e.g., science, arts, social sciences, etc.). All users performed the same retrieval task described in the previous paragraph under the three different experimental conditions. This experimental design is called "repeated measures design" (Miller, 1984), and the order in which users perform the tasks may influence their performance. For example, the task that is performed last may benefit from experience acquired in the first, or may, perhaps, suffer from the effects of fatigue or boredom. To neutralise such order effects, we varied the order in which the tasks were performed across subjects. Therefore, half the users performed first the task under condition V, whereas the other half performed first the task under condition T. Each user completed these two tasks during the same experimental session (i.e., on the same day). It was decided that all subjects should perform the task under condition C last, in a separate experimental session some time after having completed tasks V and T. This decision was based on the fact that condition C was the most complex and most difficult for the users to perform. It was our belief that if we had exposed users to condition C first, they would have been frustrated, and their performance would have been negatively biased because of the complexity of that condition. Therefore one can argue that the results for condition C reported in this paper reflect an optimistic, or an "upper bound" view of user performance. In fact, users achieved the specific results having gained experience through the other two conditions first, and it is our belief that under any other circumstances they would perform at a same or at a lower level for condition C.

5.4. Sonification of the Retrieved Document List

The experiments involved the presentation of document descriptions to subjects in three different forms, all of which were of an auditory nature. In two of the experimental conditions, the same human read the descriptions to each subject, either while physically in the same room (though not directly facing the subject), or while located in a differ-

ent room and reading the descriptions over the telephone. Care was taken not to overload the human reader, so as to avoid effects of fatigue that would bias the experimental results (e.g., no more than two sessions were performed on the same day, and there was an interval of at least 45 min between two consecutive sessions). In the third condition, a text to speech system was employed, reading the document descriptions to the users over a telephone line. The system was operated by one of the experimenters. As far as the users were concerned, they were interacting with the same system, the only difference was in the quality of the voice (human vs. speech synthesiser) and modality of access (direct vs. telephone).

User interaction with the system was defined in the following way: the system would start reading the description of the top ranked document. At any point in time, the user could stop the system and instruct it to move to the next document, or instruct it to repeat the current document description. If none of the above occurred, the system would go through the current document description, and upon reaching its end would proceed to the next description.

5.5. The Experimental Scenario

Each subject was initially briefed about the experimental process, and instructions were handed to him by the experimenter. Any questions concerning the process were answered by the experimenter. Subjects were otherwise kept ignorant of the purpose of the experiments. A set of five queries was then presented to each subject. The title and the description of each query (i.e., the "title" and "description" fields of the respective TREC topic) were read by the user, and subsequently the experimenter would start the timing for that specific query. At that point, the user would start listening to the descriptions of the retrieved documents, and would be allowed to interact with the system in one of the ways described in the previous paragraph. At all times, one of the experimenters was in the same room with the user, timing the session and overlooking the experimental process. Users had to identify relevant documents for each query within 5 min. The relevant documents were marked by the users on an answer sheet that was prepared for each query. If a user managed to examine all the documents before the specified time ended, the experimenter would record this information on the answer sheet for purposes of recording speed data. The answer sheets were returned to the experimenter after a user had finished all five queries. Once the subject had completed the assigned task in one condition, a questionnaire was handed to him. The completed questionnaire was also returned to the experimenter. The purpose of the questionnaire was to gather additional information on the user's interaction with the system, more specifically, about the utility of the document descriptions, the clarity of the voice reading the descriptions, and about the level of difficulty of the query. Therefore, the data that were collected through the above procedure from each subject comprised the answer sheets for the queries (five answer sheets per condition, one per query), and the completed questionnaires (one per condition). The analysis of the data will be presented in the following section.

TABLE 1. Average precision, recall, and time in the four experimental conditions. Conditions: on-screen display of document descriptions (S), read descriptions (V), read descriptions over the telephone (T), and finally descriptions read over the telephone by a speech synthesiser (C).

	S	V	T	C
Avg. prec. (%)	47.15	41.33	43.94	42.27
Avg. rec. (%)	64.84	60.31	52.61	49.62
Avg. time (sec)	17.64	21.55	21.69	25.84

6. Experimental Results and Analysis

In the following sections, we report the results of our experimentation. Section 6.1 describes the results, whereas section 6.2 reports an analysis of these results.

6.1. Results of the Experiments

We measured user performance in relevance assessments (the dependent variable of the experiment) in terms of accuracy and speed of the judgements. In our experiments, accuracy is defined in terms of both recall and precision. Recall represents the number of relevant documents correctly identified by a subject for a query divided by the total number of relevant documents, within the examined ones, for that query. Precision is defined as the number of relevant documents correctly identified, divided by the total number of indicated relevant documents for a query. Speed is measured in terms of time, in seconds, that a user took to assess the relevance of a single document.

Table 1 reports the results of user relevance assessments in terms of average[4] precision, recall, and time for all four experimental conditions: on-screen display of document descriptions (S), read descriptions (V), read descriptions over the telephone (T), and, finally, descriptions read over the telephone by a speech synthesiser (C).

Figures 3 and 4 present in more detail the time data collected during the experiments, by showing the average time to assess a document per user, and the average time to assess a document per condition.

Data aimed at studying the effects of fatigue in the different conditions are reported in Tables 2, 3, and 4. In Table 2, we compare the overall average time taken to assess the relevance for a document, with the average time taken to assess a document that is retrieved in response to the first and the last of the five queries making up a session. Tables 3 and 4 show analogous data for precision and recall. It should be noted that the last query was assessed after having already spent 20 min on the experimental task, and

[4]Averaged across all queries for each experimental condition.

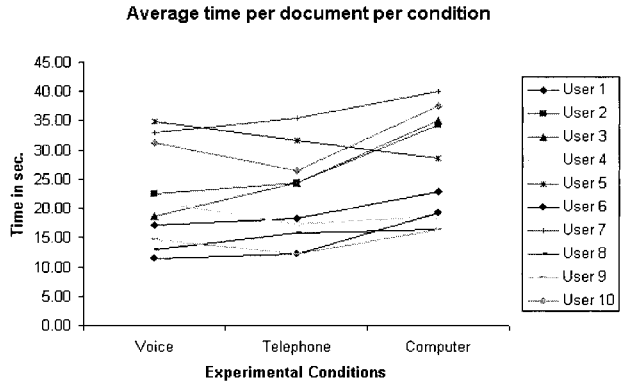

FIG. 3. Average time to assess a document per condition.

therefore the user may have been starting to lose concentration.

Table 5 shows a comparison of the effects of long and short queries on the precision, recall, and average time in the four conditions. We gathered this data to see if there was a significant variation in users' speed and accuracy in judging the relevance of documents between short queries (queries so short that users could easily keep them in mind) and long queries (queries so long that users needed to have them written down in front of them at all times and constantly refer to them). To distinguish between long and short queries, we measured the average number of lines of the description of the 50 queries used in the experiments (i.e., the part of the TREC topic that each user had to read and comprehend before starting the session). The average number of lines for the 50 queries was 5.48, and therefore we defined as "short queries" those whose descriptions contained less than six lines, and as "long queries" the remaining ones.

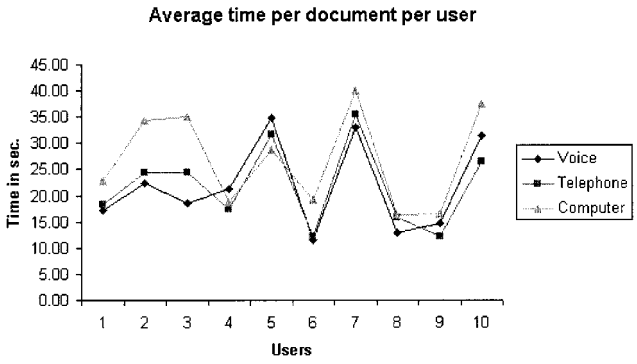

FIG. 4. Average time to assess a document per user.

TABLE 2. Average time per document per user: comparison between first and last query.

	S	V	T	C
Avg. time (sec)	17.64	21.55	21.69	25.48
Avg. time first q. (sec)	19.51	23.01	22.92	25.14
Avg. time last q. (sec)	15.26	18.41	22.27	23.38

TABLE 3. Average precision per user: comparison between first and last query.

	S	V	T	C
Avg. prec. (%)	47.15	41.33	43.94	42.27
Avg. prec. first q. (%)	40.73	57.56	48.26	57.72
Avg. prec. last q. (%)	49.25	31.39	30.21	32.41

TABLE 4. Average recall per user: comparison between first and last query.

	S	V	T	C
Avg. rec. (%)	64.84	60.31	52.61	49.62
Avg. rec. first q. (%)	59.73	65.56	43.17	48.15
Avg. rec. last q. (%)	50.85	53.06	36.81	29.33

6.2. Analysis of the Results

Table 1 shows that users in condition S perform better than any other condition in terms of precision and recall, and are also faster in their judgements. This result was expected, because condition S is the most familiar to the users and the least complex among the various experimental conditions. In this condition, the low levels of recall and precision are resulting from the difference in perception of relevance of documents between our users and the TREC assessors. Their low values should not surprise. It is a well-known fact that there is often very little agreement between two persons on the relevance of a document to a query. Comparing these recall and precision values with those obtained for other conditions shows the effects of the judgement conditions and the forms of response on the users' perception of relevance. In other words, condition S can be considered as the baseline for our analysis.

Data in Table 1 also show that performance, in terms of recall and average time, gradually decreases across conditions from S to C, although some of these differences are not statistically significant (i.e. average time of conditions V and T). A striking result is that users achieved higher

TABLE 5. Average precision, recall, and time: comparison between long and short queries.

	S	V	T	C
Long queries				
Avg. prec. (%)	48.38	43.82	49.14	39.37
Avg. rec. (%)	67.36	56.7	64.05	56.42
Avg. time (sec)	19.97	19.32	21.68	23.61
Short queries	S	V	T	C
Avg. prec. (%)	46.01	38.93	37.7	44.81
Avg. rec. (%)	64.45	63.75	43.25	44.24
Avg. time (sec)	16.51	23.70	22.69	27.12

precision in condition T than in conditions V or C. Users seem to be more concentrated when listening to the summaries over the phone than when the same summaries are read to them in person. However, the concentration did not compensate for the drop in voice quality in condition C. Nevertheless, the difference in precision between conditions S and C is not so great (only about 5%) as to create insoluble problems for a telephone-based IR system. The lower performance in terms of recall in condition C could be balanced by using relevance feedback. The correct identification of at least some relevant documents could be enough to let the relevance feedback process work effectively. This conclusion supports our intention to implement a relevance feedback mechanism in the IVIRS prototype.

A significant difference among the four conditions lies in the average time taken to assess the relevance of one document, in particular between conditions V and C (significant at the 2% level for a two-tailed *T*-test), and T and C (significant at the 5% level for a two-tailed *T*-test). This difference is large enough to enable a user to assess on average, in the same amount of time, more than 22 documents in condition S compared to only 13 in condition C, an increase of more that 70% in number of documents assessed. This result suggests that using a telephone-based IR service might be more time-consuming, and therefore more expensive, than using a conventional computer based IR system. A concerned user would have to evaluate if it is more cost effective, in terms of time connected to the service, to access the system using computer and modem and reading the documents on the screen, than accessing the system using a telephone. Of course, this consideration is only valid if the user has a choice.

An analysis of Figures 3 and 4 shows that we can conveniently divide users into two groups, depending on the speed at which they perform the relevance assessments. It should also be noted that user behaviour, as far as speed is concerned, remains consistent across all three experimental conditions. In other words, "fast" users remain fast, and "slow" users remain slow, whatever the experimental condition. Figure 3 best represents this observation: one can almost perfectly divide the set of slow and fast users into two classes by drawing a horizontal line that defines the time point (at approximately 17.5 sec) that distinguishes the two groups. The hypothesis that slow users are more accurate in their judgements was not proved by our data.

Table 2 shows that in all experimental conditions the average time per document is lower for the last query than for the first one (significant at the 2% level for a two-tailed *T*-test for condition V). Moreover, Tables 3 and 4 show that both precision and recall values for the last query are significantly lower that those for the first query (significant at the 5% level for a two-tailed *T*-test for both precision and recall in condition C). We believe that this result indicates that users cannot hold their concentration on the telephone for a long period of time. It seems to be the case that, after some time, users start making hasty and often erroneous judgements. Therefore, it may be more effective for a user to have many short sessions with a telephone-based IR service instead of a single long one.

Although the data in Table 5 do not statistically confirm any findings, we can observe that users tend to be faster and more precise with long queries than with short ones. The only exception to this is condition C, where precision was higher with short queries than with long ones. A possible explanation for these results can be given by examining how the user marked the descriptions of the queries on the answer sheets. When presented with long queries, users tended to mark a few "key words" in the description of the query, and subsequently look for these key words in the document surrogates. However, users did not usually follow the same technique when examining short queries. This technique of identifying key words seems to enable users to identify more precisely relevant documents in conditions S, V, and T, but does not seem to work for condition C. A possible explanation is that, in condition C, users might not be able to spot the key words because of the poor quality of the synthesised voice. This would explain the considerable drop in precision for long queries from condition T to C. With short queries, users did not usually mark key words and concentrated more on listening to the document descriptions. Nevertheless, short queries were more difficult to assess than long ones because of their ambiguity. It is probably the further increase in attention necessary to deal with condition C that explains the much longer average time and the higher precision of users dealing with short queries in this condition compared with other conditions. Finally, an analysis of the data collected through the questionnaires showed that there was no great difference in perception of query complexity and usefulness of the document descriptions among conditions. Most users found the voice of the human reader clear, as opposed to the voice of the speech synthesiser, which they found hard to understand and tiring to listen to for a long time.

7. Conclusions and Future Work

We presented the results of a study of users' perception of relevance of documents. Documents retrieved in response to a query are presented to the users in a variety of conditions and we compared the differences in users' perception of relevance related to the judgement conditions and forms of response (Cuadra & Katter, 1967). Our results suggest that users' perception of relevance of documents is highly influenced by these factors. In the particular case of spoken documents, the low levels of accuracy and speed of the judgements suggest the necessity of studying more sophisticated ways of presenting documents to users and more complex forms of human–computer interaction.

The most important implications of our results for the design and implementation of the IVIRS system, and of similar systems, are the following:

- The system should enable the user to provide interactive relevance feedback to the retrieval process, because this

would increase the performance of the system as perceived by the user.
- The system should provide to the user during the query session an indication of the actual cost of the service. Given the increase in the average time necessary to assess a document, a concerned user would then be able to evaluate at any stage of the interaction whether the service is cost effective or not.
- The system should be designed to handle short sessions with the user, because this seems to be the most effective method of use. For example, the system should retrieve and present only a small number of documents in response to a query, because this will avoid tiring the user and leading him to make inaccurate judgements.

Finally, more studies on the effects of voice synthesis, intonation and speed are necessary, as well as the design of new techniques to produce document summaries targeted at speech interaction. In the context of the SIRE project we are currently experimenting with such issues. A prototype of the system is currently under way. The prototype will provide a useful experimental tool for research on the sonification of an IR environment.

Acknowledgments

We would like to thank Jane Reid for her help during the experiments (she was the human reader) and for her useful comments on a draft version of this paper. We are also grateful to the reviewers for their helpful suggestions.

References

Arons, B. (1997). Speech skimmer: A system for interactively skimming recorded speech. ACM Transactions on Computer–Human Interaction, 4, 3–38.

Barry, C.L. (1995). User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3), 149–159.

Bernsen, N., Dybkjoer, H., & Dybkjoer, L. (1997). What should your speech system say? IEEE Computer, pp. 25–31.

Brandow, R., Mitze, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing and Management, 31, 675–685.

Crestani, F. (1998). Sonification of an information retrieval environment: Design issues. In International forum on multimedia and image processing. (pp. 789–794) Anchorage, AL.

Crestani, F. (1999). Vocal access to a newspaper archive: Design issues and preliminary investigations. In Proceedings of ACM Digital Libraries (pp. 59–66). Berkeley, CA.

Cuadra, C., & Katter, R. (1967). Opening the black box of relevance. Journal of Documentation, 23, 291–303.

Edmundson, H. (1964). Problems in automatic abstracting. Communications of the ACM, 7, 259–263.

Goose, S., Wynblatt, M., & Mollenhauer, H. (1998). 1-800-hypertext: Browsing hypertext with a telephone. In Proceedings of ACM Hypertext (pp. 287–288). Pittsburgh, PA.

Harman, D. (1996). Overview of the fifth text retrieval conference (TREC-5). In Proceedings of the TREC Conference, Gaithersburg, MD.

Keller, E. (Ed.) (1994). Fundamentals of speech synthesis and speech recognition. Chichester, UK: John Wiley and Sons.

Luhn, H. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159–165.

Maizell, R., Smith, J., & Singer, T. (1971). Abstracting scientific and technical literature: An introductory guide and text for scientists, abstractors and management. New York: Wiley-Interscience, John Wiley and Sons.

Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M., & Sundheim, B. (1998). The TIPSTER SUMMAC text summarization evaluation: Final report. MITRE Corporation Technical Report.

Miller, S. (1984). Experimental design and statistics, 2nd ed. London, UK: Routledge.

Mittendorf, E., & Schauble, P. (1996). Measuring the effects of data corruption on information retrieval. In Proceedings of the SDAIR 96 Conference (pp. 179–189). Las Vegas, NV.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48, 810–832.

Paice, C. (1990). Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management, 26, 171–186.

Peckham, J. (1991). Speech understanding and dialogue over the telephone: An overview of the ESPRIT SUNDIAL project. In Proceedings of the Workshop on Speech and Natural Language (pp. 14–27). Pacific Grove, CA.

Rush, J., Salvador, R., & Zamora, A. (1971). Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. Journal of the American Society for Information Science, 22, 260–274.

Sanderson, M., & Crestani, F. (1998). Mixing and merging for spoken document retrieval. In Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (pp. 397–407). Crete, Greece.

Saracevic, T. (1970). The concept of "relevance" in information science: A historical review. In Saracevic, T. (Ed.), Introduction to Informatiuon Science, 111–151. R.R. Bowker, New York, USA.

Schamber, L., Eisenberg, M.B., and Nilan, M.S.(1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26, 755–776.

Sparck-Jones, K., Jones, G., Foote, J., & Young, S. (1996). Experiments in spoken document retrieval. Information Processing and Management, 32, 399–417.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In Proceedings of ACM SIGIR (pp. 2–10). Melbourne, Australia.

Vickery, B.C. (1959). Subject analysis for information retrieval. Proceedings of the International Conference on Scientific Information, 2, 855–865.

Voorhees, E., Garofolo, J., & Spark-Jones, K. (1997). The TREC-6 spoken document retrieval track. In TREC-6 notebook (pp. 167–170). Gaithersburg, MD: NIST.