INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. ● Suite 600 ● Berkeley, California 94704-1198 ● (510) 643-9153 ● FAX (510) 643-7684

A Study of Users' Perception of Relevance of Spoken Documents

Tassos Tombros* **Fabio Crestani†**

TR-99-013

July 1999

Abstract

We present the results of a study of users' perception of relevance of documents. 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 aim is to study experimentally how users' perception of relevance varies depending on the form that retrieved documents are presented. 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.

*Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, UK, email: tombrosa@dcs.gla.ac.uk

†International Computer Science Institute, 1947 Center St. Suite 600, Berkeley, CA 94704, USA, email: fabioc@icsi.berkeley.edu

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 increasing emphasis on extending the utility of information systems, by providing access to them through mobile devices, for example telephones or PDAs [10]. Enabling access to an Information Retrieval (IR) 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 user community. In addition to offering greater convenience and flexibility, ubiquitous access to IR 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 IR services without any of the problems encountered using a computer.

The retrieval of spoken documents using a textual query is also a fast emerging area of research (see for example [23]). It involves an efficient, more 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 retrieval track in the TREC-6 conference [26]. The problem is to devise IR models that can cope with the large number of errors inevitably found in the transcripts of the spoken documents. Models designed for retrieval of OCRed documents have proved useful in this context [17].

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. While the incorrect or uncertain recognition of an instance of a word in a long spoken document can be compensated 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 the failing of recognising a query word, or worse, the incorrect recognition of a query word will fail to retrieve a large number of relevant documents or wrongly retrieve a large number of non-relevant 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 will 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 will find it difficult to understand the response of the system and will not be able to use it as efficiently as a conventional on-screen IR system. These difficulties need to be addressed to be

[1] There is an on-going debate about realistic query lengths. While TREC queries are on average about 40 words long, Web queries are only 2 words long on average. This recently motivated the creation in TREC of a "short query" track, to experiment with queries of more realistic length.

able to implement such a system effectively.

This paper is concerned with the last of the above issues: evaluating the possibility of effective use of a telephone based IR service from the user's perspective. In particular we addressed and test 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 paper is structured as follows. Section 2 describes the background and the motivations of this study, putting it in a wider context, while 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. Section 4.1 describes the document summarisation tool we used in our experimentation. The experimental design of our user study is reported in Section 5, and the results are described and analysed in Section 6. Section 7 reports the conclusions of our work and points at directions of future extensions of the work reported in this paper.

2 Background

The background of the work reported in this paper is related to a project currently under way at the University of Glasgow: the SIRE project. The main objective of the project is to enable a user to interact (e.g., submit queries, commands, relevance assessments, and receive summaries of retrieved documents) with a probabilistic IR system over a low bandwidth communication line (e.g., a telephone line). The next two sections describe the overall goal of the SIRE project and how the study reported in this paper fits into the it.

2.1 The SIRE Project

In the second half of 1997, at Glasgow University, we started a small project on the sonification of an IR environment. The project is funded by the European Union under the Training and Mobility of Researchers (TMR) scheme of the European Union Fourth Framework of projects. The main objective of the project is to enable a user to interact (e.g., submit queries, commands, relevance assessments, and receive summaries of retrieved documents) with a probabilistic IR system over a low bandwidth communication line, like for example a telephone line. An outline of the system specification of the prototype being built is reported in figure 1.

The prototype *interactive vocal information retrieval system* (IVIRS), resulting from the "sonification" of a probabilistic IR system, is made up of the following components:

- a *vocal dialog manager* (VDM) that provides an "intelligent" speech interface between user and IR system;

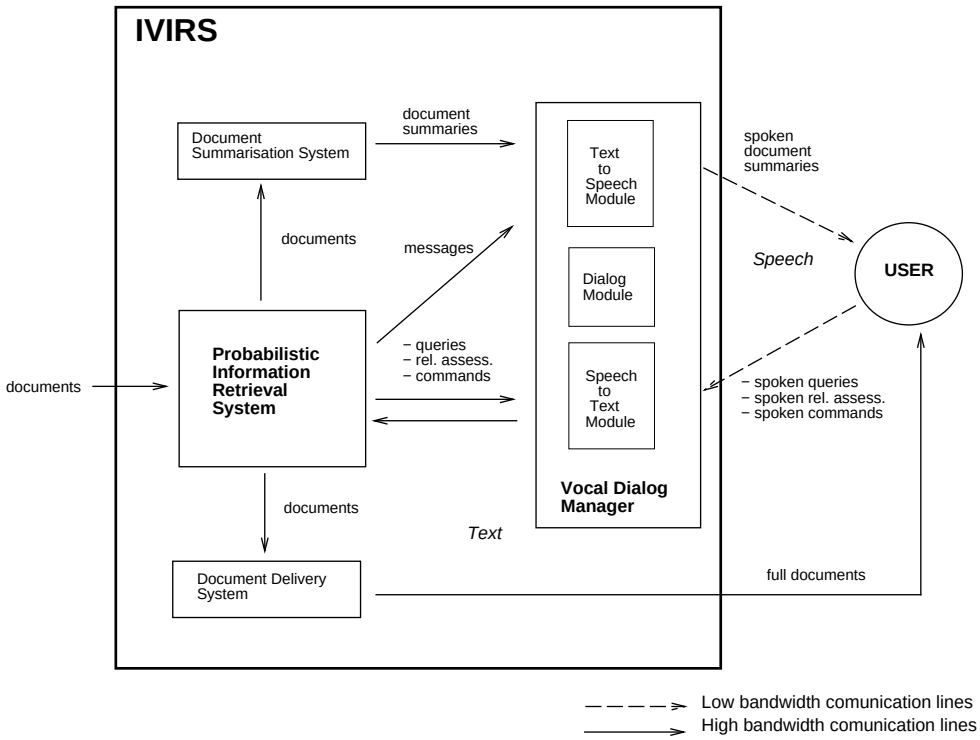

Figure 1: Schematic view of the IVIRS prototype

- 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 a summary 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 recognised the user by means of a username and a password (to avoid problems in this phase we devised a login procedure based on keying in an identification number using a touch tone), the user submit a spoken query to the system. The VDM interact 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 in order to limit the effects of wrongly recognised words in the query, Additionally, 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 case of a uncertain recognition of a word, asking to re-utter the 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 a set of document regarded as very likely to be relevant (this feature can be set in the most appropriate way by the user). The user is informed on 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 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 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 all together. Marked documents are stored in a retrieved set and the user can proceed with a new query if he wishes so. 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 case of wrongly recognised query words, since the confidence values of query words increase 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 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 already 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, while others are new to us. Some components were found not to be fully suitable to the task and had to be developed. This is the case of the probabilistic IR system, and a 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 [7]. A "divide et impera" approach has been followed, consisting of dividing the implementation and experimentation of IVIRS in the parallel implementation and experimentation of different 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 [22], and the VDM [6].

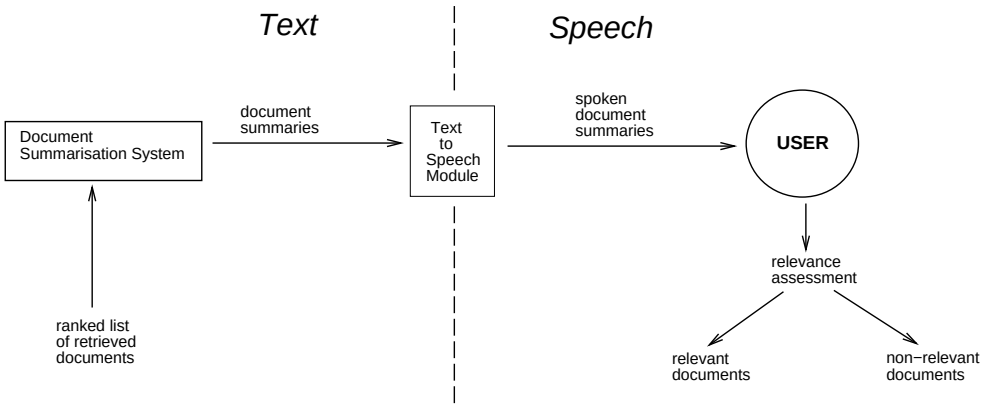

Figure 2: Generation of spoken document summaries of retrieved documents.

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 essential for an effective use of the system, since the identification of relevant documents could trigger a relevance feedback process that would not be efficient if fed with non relevant documents.

However, results of investigations in other application areas (see for example [20, 2]) showed that the previous 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 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 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 towards satisfying his information need. The goal of an IR system is to retrieve in response to a query all and only the relevant documents known to the IR system. 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.

Despite the fact that the concept of relevance is central to IR, and despite numerous research attempts to precisely define it, a satisfactory definition has not yet been given [18]. Although in the mathematical and probabilistic literature one can find many references to the term relevance, in [4] for example there is an entire chapter devoted to "Relevance and Irrelevance", there is no explicit theoretical notion of relevance that can be used in the context of IR.

Currently, there are two main views of relevance in IR:

- *topic-appropriateness*, 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, despite the fact that the above mentioned distinction is widely accepted.

Topic-appropriateness is related to a so called "system-perceived relevance", since it does not involve any user judgement, and it is left completely to the IR system. Because of this, topic-appropriateness can be considered "objective" and can be studied only from the point of view of the IR system and the IR process. This view of relevance has found many supporters among IR researchers, for example Cooper [5].

On the other hand, the user-utility view of relevance has a much broader sense than topic-appropriateness, and it involves a much deeper knowledge of the user information need and of the purpose of this need. We can relate this view to a so called "user-perceived relevance", where the relevance of a document to an information need is left completely to the user's judgement. The user-utility notion of relevance is therefore a "subjective" notion: different users may have radically different views about the relevance or non-relevance of particular documents to an information need, since they may perceive it a completely different ways. This view has also found many supporters, for example Van Rijsbergen [25, pp.146].

In this paper we are mainly concerned with the second notion of relevance. A user-utility notion of relevance assumes that a user is able to "understand" whether a document is relevant to his information need or not. In fact, we are interested in evaluating how the user's perception of a document's relevance 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 [8] that could be grouped into six classes: people, documents, statements of information

requirements, judgement conditions, form of response, and judgemental attitudes. In the work reported in this paper, 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, in 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 multi-modal 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 relevance judgements 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 the most important 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 [1]. 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 was 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 [14]. Furthermore, 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 focussed towards this particular information need than a generic, non query sensitive summary. In this way summaries can provide an indicative function, serving as a preview format to support relevance assessments on the full text of documents [21].

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 [24]. 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 [15] 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 [19] 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 [24]; here we shall briefly describe the summary generation process. The document collection to be summarised comprised news articles of the Wall Street Journal (WSJ) taken from the TREC collection [11]. 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) in order 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 [13]. 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 [9, 3]. The summary produced by the system can be presented to the user by means of a very simple interface designed as a web page (i.e., written in HTML) and that can be displayed using a common web browser. Figure 3 reports some examples of such 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. Often, this takes the form of a text-to-speech system, whereby unrestricted text is transformed into speech. Since most on-line information is represented as ASCII text, the automatic conversion of text to speech provides a means to present many people with on-line 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 [12]. We carried out a survey and an initial testing of a number of commercially available speech synthesis systems. After 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, and not of the quality available only on experimental research prototype system. For the experiments reported in this paper we used the Monologue 97 system[2]. 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 since it enables to make adjustments for a variety of voice characteristics.

However, given the limits and the quality of state-of-the-art speech synthesis sys-

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

Query No. 08

You have just typed in the following query:

"Alternatives to Postscript"

To be relevant, a document must identify a need for, or the existence of, an alternative to Postscript, a page description language.

The following articles have been retrieved in relation to that query:

Documents 1-10 (of 50) matching the query.

[1]

Publishing: Papers Take Alternative Path to Success ---- By Andrew Patner Staff Reporter of The Wall Street Journal

As publisher of the Chicago Reader, an alternative weekly newspaper, Robert A. Roth refuses to read daily papers. He's afraid their stodgy ways might be infectious. The 68 papers that belong to the Association of Alternative Newsweeklies now reap $100 million-plus in combined revenue, and their 3.5 million total circulation includes an enviable share of the coveted 18- to 35-year-old market, the very group that daily newspapers are having the most trouble attracting. The alternative weeklies group admitted seven new papers this year and had applications from seven more. Cartoonists Lynda J. Barry, whose work appears in Esquire, Mother Jones and other magazines, and Matt Groening, of "The Simpsons," started with alternative papers and still do weekly strips for them.

[Click here to get the whole article]

[2]

School Days: Break the Teaching Monopoly ---- By C. Emily Feistritzer

There are vast numbers of adults with at least a bachelor's degree who want to teach. Many have advanced degrees and years of successful experience in other careers. That means one has to go to college and complete a series of education courses approved by the state department of education. Alternative teacher certification programs are not wanting for detractors. Since last week, after an article about our study of alternative teacher certification appeared in the New York Times, we've gotten scores of phone calls from people who want to know where the alternative certification programs are -- they included a dentist who wants to teach biology, an international businessman with a master's degree in physics who wants to teach high school physics and math, and a corporate executive who wants to teach elementary school.

[Click here to get the whole article]

Figure 3: Example of document summaries.

tems, we decided to introduce in our experiments with 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 in order to simulate degrading levels of the quality of speech. The details of the experimental procedure are reported in the following section.

5 Experimental Design

In this section we describe the experimental design of our investigation, providing the details of the experimental conditions, the tasks, the subject grouping, and the experimental scenario.

Experimental conditions

The aim of the experiments reported in this paper is to investigate the effects of different forms of presentation of document descriptions on users' perception of document relevance. In a previous study Tombros and Sanderson used document titles, and automatically generated, query-biased summaries as document descriptions, and measured user ability to make fast and accurate relevance judgements when the descriptions were displayed on a computer screen and read by the users [24]. 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 also show that any variation in user performance between the experimental conditions can be attributed only to changes in the independent variable, since 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 [16].

Task

In order to be able to use the experimental results reported in [24], 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 minutes. The information that was 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), set of retrieved documents for each query (the 50 top-ranked documents were presented to each user), and document descriptions (titles, and the query-biased summaries) as in [24]. Queries were randomly allocated to subjects by means of a draw, but since each subject was presented with a total of 15 queries (5 queries for each condition) we ensured that no query was assigned to a specific user more than once.

Groups of subjects

A group consisting of 10 users was employed. The population was drawn from post-graduate students doing a conversion course in information technology. All users performed the same retrieval task described in the previous paragraph under the three different experimental conditions (repeated measures design [16]). The order in which users perform the tasks in this experimental design 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. In order to neutralise such order effects, we had to vary the order in which the tasks were performed across subjects. Therefore, half of 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 due to the complexity of that condition.

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 different 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 minutes between two consecutive sessions). In the third condition, a text to speech system was employed[3] 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

[3]We carried out a test of a number of text to speech systems and we chose First Byte's Monologue'97, because of the quality of the voice and ease of use.

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.

The experimental scenario

Each subject was initially briefed about the experimental process, and instructions were handed to him/her 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 5 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 minutes. 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. 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 anyone condition, a questionnaire was handed to him/her. 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, and the completed questionnaires. The analysis of the data will be presented in the following section.

6 Experimental Results and Analysis

In the following sections we report the results of our experimentation. Section 6.1 describes the results, while 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. 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 takes to assess the relevance of a single document.

Table 1 reports the results of user relevance assessments in terms of average 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).

	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	25.48

Table 1: Average precision, recall, and time in the four experimental conditions.

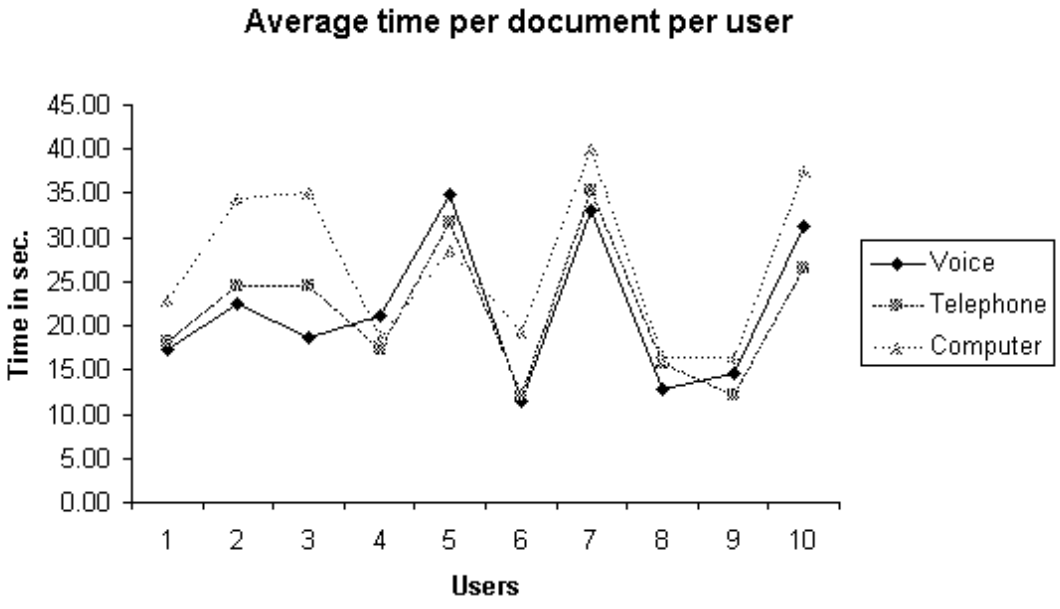

Figure 4: Average time to assess a document per user.

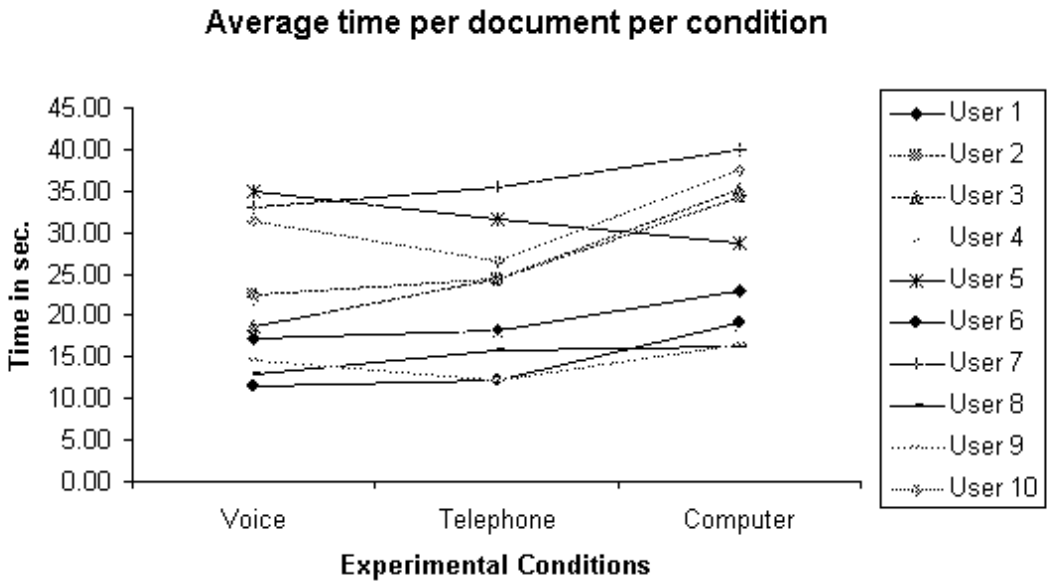

Figure 5: Average time to assess a document per condition.

Figures 4 and 5 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.

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

Data aimed at studying the effects of fatigue in the different conditions, are reported in tables 2, 3 and 4. In tables 2 we compare the overall average time taken to assess the relevance for a document, with the average time required to assess a document that is retrieved in response to the first and the last of the 5 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 minutes on the experimental task and therefore the user was surely starting loosing 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 in order to see if there was a significant variation in user's speed and accuracy in judging the relevance

	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 3: Average precision 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

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

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 from of them at all time). In order 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 6 lines, and as "long queries" the remaining ones.

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, since 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. Comparing these recall and precision values with those obtained for other conditions shows the effects of the judgement conditions and the forms of response in the users' perception of relevance. In other words condition S can be considered as the baseline for our analysis.

Recall and average time gradually decrease across conditions from S to C, although some of these differences are not statistically significant (i.e., average time of V and T). A striking result is that users achieved higher precision in condition T than in conditions V or C. Users seem to be more concentrated when listening to the summaries over the phone. 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

Long queries	S	V	T	C
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

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

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 S and C. 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. However, this may not be possible for visually impaired users.

An analysis of figures 4 and 5 shows we can conveniently divide users into two groups, depending on the speed by 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 5 best represents this observation: one can almost perfectly divide the set of 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. 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. 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 one long one.

Although there is not enough data in table 5 to strongly 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 "keywords" in the description of the query, and subsequently look for these keywords in the description of the documents. However, users did not usually follow the same technique when examining short queries. This technique of identifying keywords seems to enable users to identify more precisely relevant documents in conditions S, V and T. However, this technique does not seem to work for condition C. A possible explanation is that in condition C users might not be able to spot the keywords 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 keywords and concentrated more on listening to the document descriptions. Nevertheless, short queries were more difficult to assess than long ones due to 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 to be hard to understand and tiring to listen 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 [8]. 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 in 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 relevance feedback to the retrieval process, since this would increase the recall 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. A concerned user would then be able to evaluate if the service is cost effective compared to a conventional on screen service.
- The system should be designed to handle short sessions with the user, since this seems to be the most effective way of using it. For example, the system should not retrieve and present many documents in response to a query, because this would tire the user and lead him to 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.

Acknowledgements

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.

References

[1] B. Arons. SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction*, 4(1):3–38, 1997.

[2] N.O. Bernsen, H. Dybkjoer, and L. Dybkjoer. What should your speech system say? *IEEE Computer*, pages 25–31, December 1997.

[3] R. Brandow, K. Mitze, and L.F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[4] R. Carnap. *Logical Foundations of probability.* Routledge and Kegan Paul Ltd, London, UK, 1950.

[5] W.S. Cooper. A definition of relevance for Information Retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[6] F. Crestani. Sonification of an Information Retrieval environment: design issues. In *International Forum on Multimedia and Image Processing*, Anchorage, Alaska, USA, May 1998.

[7] F. Crestani. Vocal access to a newspaper archive: design issues and preliminary investigations. In *Proceedings of ACM Digital Libraries*, Berkeley, CA, USA, August 1999. In press.

[8] C.A. Cuadra and R.V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4):291–303, 1967.

[9] H.P. Edmundson. Problems in automatic abstracting. *Communications of the ACM*, 7(4):259–263, 1964.

[10] S. Goose, M. Wynblatt, and H. Mollenhauer. 1-800-hypertext: browsing hypertext with a telephone. In *Proccedings of ACM Hypertext*, pages 287–288, 1998.

[11] D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *Proceedings of the TREC Conference*, Gaithersburg, MD, USA, November 1996.

[12] E. Keller, editor. *Fundamentals of Speech Synthesis and Speech Recognition.* John Wiley and Sons, Chichester, UK, 1994.

[13] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, pages 159–165, April 1958.

[14] R.E. Maizell, J.F. Smith, and T.E.R. Singer. *Abstracting scientific and technical literature: An introductory guide and text for scientists, abstractors and management.* Willey-Interscience, John Willey and Sons Inc., New York, USA, 1971.

[15] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. The TIPSTER SUMMAC text summarization evaluation: final report. Unpublished paper, September 1998.

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

[17] E. Mittendorf and P. Schauble. Measuring the effects of data corruption on Information Retrieval. In *Proceedings of the SDAIR 96 Conference*, pages 179–189, Las Vegas, NV, USA, April 1996.

[18] S. Mizzaro. Relevance: the whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[19] C.D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[20] J. Peckham. Speech understanding and dialogue over the telephone: an overview of the ESPRIT SUNDIAL project. In *Proceedings of the Workshop on Speech and Natural Language*, pages 14–27, Pacific Groce, CA, USA, February 1991.

[21] J.E. Rush, R. Salvador, and A. Zamora. 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(4):260–274, 1971.

[22] M. Sanderson and F. Crestani. Mixing and merging for spoken document retrieval. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 397–407, Crete, Greece, September 1998.

[23] K. Sparck Jones, G.J.F. Jones, J.T. Foote, and S.J. Young. Experiments in spoken document retrieval. *Information Processing and Management*, 32(4):399–417, 1996.

[24] A. Tombros and M. Sanderson. Advantages of query biased summaries in Information Retrieval. In *Proceedings of ACM SIGIR*, pages 2–10, Melbourne, Australia, August 1998.

[25] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[26] E. Voorhees, J. Garofolo, and K. Spark Jones. The TREC-6 spoken document retrieval track. In *TREC-6 notebook*, pages 167–170. NIST, Gaithersburgh, MD,USA, 1997.