Vocal Access to a Newspaper Archive: Assessing the Limitations of Current Voice Information Access Technology

Fabio Crestani
Department of Computer Science
University of Strathclyde
26 Richmond Street
Glasgow G1 1XH, Scotland, UK
Email: F.Crestani@cs.strath.ac.uk

10th July 2002

Abstract

This paper presents the design and the current prototype implementation of an interactive vocal Information Retrieval system that can be used to access articles of a large newspaper archive using a telephone. The implementation of the system highlights the limitations of current voice information retrieval technology, in particular of speech recognition and synthesis. We present our evaluation of these limitations and address the feasibility of intelligent interactive vocal information access systems.

Keywords: Information Retrieval, Speech Applications, Intelligent Interactive Information Access Systems.

1 Introduction

For the last 50 years Information Retrieval (IR) has been concerned with enabling users to retrieve textual documents (and for a long time only bibliographic references) in response to textual queries. With the availability of faster computers and cheaper electronic storage it is now possible to store and search the full text of millions of textual documents online.

It has been observed that widespread access to Internet and the high bandwidth often available to users makes them assume that there is nothing easier than connecting to and searching a large information repository. Nevertheless, a very large part of the world population does not and will not have for a long future access to computers or Internet. Moreover, there are cases, even in developed countries, in which users may need to access information without using a computer screen or keyboard and without Internet connection. Indeed, there are cases in which the most convenient communication mean is a telephone, especially a mobile phone. We could provide a number of examples of such situations: an automatic booking service for an airline company, a press archive for journalists, an automatic telephone based customer support service, a travel or weather advisor, in-car applications, and many ubiquitous computing applications. In all these cases it is necessary to have access and interact with a information system solely via voice. This implies the design and implementation of systems capable not only of understanding the user's spoken request, finding the required information and presenting it as speech, but also capable of interacting with the user in order to better understand the user's information need, whenever this is not clear enough to proceed effectively with the searching. In addition, there are also situations and users for whom the availability of computer screen and keyboard is not useful. Blind or partially-sighted users (e.g. users who have problems due to disabilities, protective clothing or working environment) may only have access to information if this is accessible via voice. This paper is concerned with the design, feasibility study, and implementation issues of one such voice information access systems.

The paper is organised as follows. Section 2 introduces the difficult union between IR and speech. Section 3 gives the background of this work and explains its final objective: an Interactive Vocal Information Retrieval System. Section 4 reports on the current state of the implementation of the first prototype system. Section 5, the core of the paper, reports the results of a feasibility study that addresses some of the limitations of current speech technology (mainly speech recognition and synthesis) for the implementation of effective interactive vocal information access systems. The question we want to answer is: does current speech technology enable to effectively implement such systems? Section 6 reports on work in progress toward overcoming some of these limitations. Section 7

summarises the conclusions of our study.

2 Information Retrieval and Speech

Information Retrieval (IR) is the branch of computing science that aims at storing and allowing fast access to a large amount of multimedia information, like for example text, images, speech, and so on [31]. An *Information Retrieval System* is a computing tool that enables a user to access information by its semantic content using advanced statistical, probabilistic, and/or linguistic techniques.

Most current IR systems enable fast and effective retrieval of textual information or documents, in collections of very large size, sometimes containing millions of documents. The retrieval of multimedia information, on the other hand, is still an open problem. Very few IR systems capable of retrieving multimedia information by its semantic content have been developed. Often multimedia information is retrieved by means of an attached textual description.

The marriage between IR and speech is a very recent event. IR has been concerned for the last 50 years with textual documents and queries. It is only recently that talking about multimedia IR has become possible. Progress in speech recognition and synthesis [15, 11] and the availability of cheap storage and processing power have made possible what only a few years ago was unthinkable.

Since the two main types of objects and IR system deals with are queries and documents, the association between IR and speech has different possibilities:

- textual queries and spoken documents;
- spoken queries and textual documents;
- spoken queries and spoken documents.

It should be noted that in the large majority of current IR systems capable of dealing with speech, the spoken documents or queries are first transformed into their textual transcripts and then dealt by the IR system with techniques that are derived from those used in normal textual IR.

The retrieval of spoken documents using a textual query is a fast emerging area of research (see [1] for a recent overview). It involves the combination of the most advanced techniques of speech recognition and IR. The increasing interest in this area of research is confirmed by the inclusion, for the first time, of a spoken documents retrieval (SDR) track in the TREC-6 conference [32]. The

problem here 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 [20]. Another problem is related to the fact that, although IR models can easily cope with the fast searching of large document collections, fast speech recognition of a large number of long spoken documents is a much more difficult task. Because of this, spoken documents are converted into textual transcripts off-line, and only the transcripts are dealt by the IR system.

The problem of retrieving textual documents using a spoken query may seem easier than the previous one, 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 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 and wrongly retrieve a large number of non-relevant documents. Nevertheless, spoken query processing (SQP) has not received the same level of attention that SDR.

The retrieval of spoken queries in response to spoken documents is a very complex task and is more in the realm of speech processing than IR, although IR techniques could be useful. Speech recognition and processing techniques could be used to compare spoken words and sentences in their raw form, without the need of generating textual transcripts. However, this approach can only work in very restrictive domains and environments, with a very small vocabulary, and with a very restricted number of users.

3 An Interactive Vocal Information Retrieval System

The main objective of the SIRE (Sonification of Information Access) 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 system, like for example a telephone line. An outline of the system specification is reported in the figure 1.

This *interactive vocal information retrieval system* (IVIRS), resulting from the

[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.

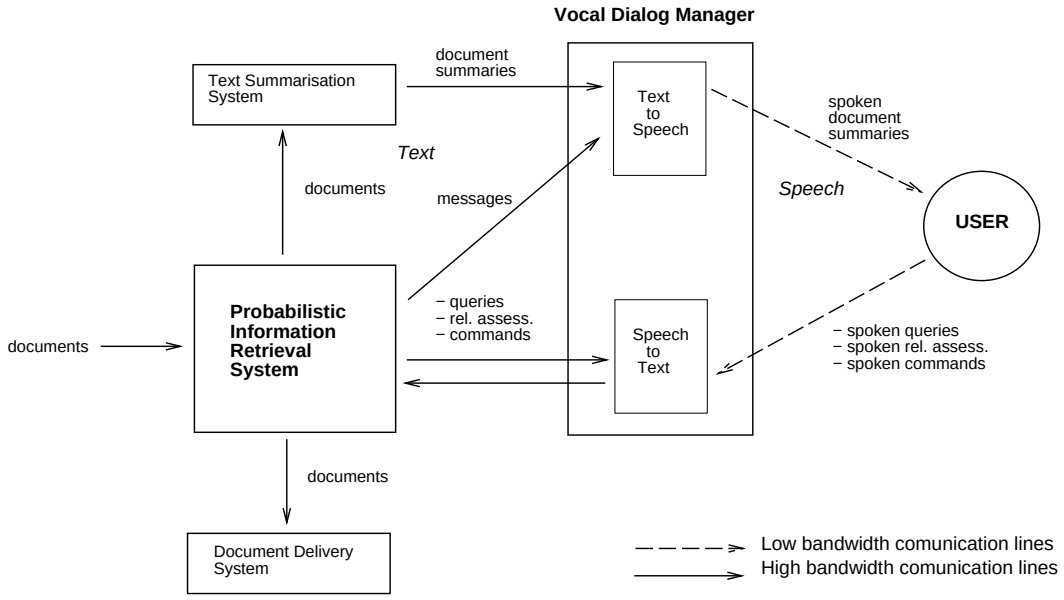

Figure 1: Schematic view of the IVIRS prototype

"sonification" of a probabilistic IR system, has the following components:

- a *vocal dialogue 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 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.

IVIRS 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 speaker identification 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 the 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 can also fed to the PIRS. This information can be useful to limit the effects of wrongly recognised words in the

query. Additionally, an effective interaction between the system and the user often helps to tackle this problem. In fact, in case of problems in the recognition of an important word, the system asks the user 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; a threshold is used to find the a set of document regarded as surely relevant. 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 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 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 a known address by the DDS.

3.1 The Vocal Dialogue Manager

The human-computer interaction performed by the *VDM* is not a simple process that can be done by off the shelf devices. The VDM needs to interact in an "intelligent" way with the user and with the PIRS in order to understand completely and execute correctly the commands given by the user. In fact, while the technology available to develop the speech synthesis (Text-to-Speech) module is sufficient for this project (but see section 5.1), the technology available for the speech recognition (Speech-to-Text) module is definitely not. On one hand, the translation of speech into text is a much harder task than the translation of text to speech, in particular in the case of continuous speech, speaker independent, large vocabulary, and noisy channel speech recognition. On the other hand, it is necessary to take into consideration the three possible forms of uncertainty that will be present in the IR environment when we add a vocal/speech component:

1. the uncertainty related to the speech recognition process;
2. the uncertainty given by the word sense ambiguity present in the natural language;
3. the uncertainty related to the use of the spoken query in the retrieval process.

In order to deal with these different forms of uncertainty we need not only to developed an advanced retrieval model for the PIRS, but also develop a model of the user-VMD-PIRS interaction that encompasses a language model, an interaction model and a translation model. This later part of the project is still at the early stages of development and will not be presented here. In this context we also make use of the results of previous work in this area, although in rather different applications (see for example [22, 3, 18]).

Currently we are building a model of the telephone interactions between users and PIRS analysing the results of a study carried out using the "Wizard of Oz" technique. The Wizard of Oz technique is a way of simulating human-computer interfaces. Unknown to the user, a human "wizard" performs some of the functions of the computer such as responding to the user's spoken input. The technique is commonly used where the technology for running the interface does not yet exist or is not yet sophisticated enough to be used in real time. The particular Wizard of Oz simulation we are currently using for the design of the VDM incorporates a simple statistical model of word recognition errors based on a large sample of recognised text, whereby a realistic distribution of speech recogniser errors can be generated at any desired overall accuracy level. It enable a human behind the interface to simulate recognition errors and generate incorrect responses.

A limitation to the realism of this form of simulation is that recognition performance depends only on the content of the user's input, and not on its quality (clarity of speaking, background noise etc) as it would with a real recogniser. This is particularly relevant in the case of word-spotting, where the recogniser is designed to pick out instances of keywords embedded in arbitrary speech. Typically the accuracy of a real word-spotter is better for isolated keyword utterances than for embedded ones. To address this, a second-generation simulation method is currently being developed, in which the wizard's input (giving the keyword content of the utterance) is combined with acoustic information extracted automatically from the speech signal. This will enable us to design appropriate error recovery strategies whenever one or more words in the spoken query are below a certain threshold of recognition. A detailed presentation of this work in progress is outside the scope of this paper.

The VMD has two sub-components: a Speech-to-Text module and a Text-to-Speech module.

The *Speech-to-Text module* is arguably the most important and the most problematic module of the VMD. Speech recognition has been progressing very rapidly in the last few years [15] and results are improving day by day, in particular for speaker dependent systems. There is already a number of system commercially available that guarantee quite impressive performance once they have been prop-

erly trained by the user (e.g. Dragon Naturally Speaking, IBM Via Voce, etc.). The user also needs to teach how to recognise words that are not part of the recogniser's vocabulary. The situation is not so good with speaker independent continuous speech recognition systems over the telephone, although a number of studies have achieved some acceptable results [16, 21]. In SIRE we do not intend to develop any new speech recognition system. This is a difficult area of research for which we do not have necessary know-how. Instead, we make use of available state-of-the-art technology and try to integrate it in our system. The aim is to enable the PIRS to deal with the uncertainty related to the spoken query recognition process, integrating this uncertainty with the classical uncertainty intrinsic in the IR process.

The *Text-to-Speech module* of the VDM uses the state-of-the-art technology in speech synthesis. We carried out a survey and an initial testing of available commercial speech synthesis systems and we found that despite recent advances, speech synthesis technology is still limited. The quality of the synthesised voice is often poor and difficult to understand for a casual user. There is a large number of applications that make use of speech synthesis (see for example [10]), but their scope is limited to domains where the user could somewhat guess what the systems say. Only very few studies are available on the effectiveness of speech synthesis in very broad application domains [11]. For the experiments reported in this paper we used the Monologue 97 system[2]. Monologue 97 uses the PrimoVox Speech Synthesiser 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. We are currently upgrading the VDM to use a more advanced text-to-speech software, since important progress has been made in speech synthesis is the last few years.

Both the Speech-to-Text module and the Text-to-Speech module are implemented in VoiceXML.

3.2 The Probabilistic Information Retrieval System

The *PIRS* performs a ranking of the document in the collection in decreasing order of their estimated probability of relevance to the user's query. Past and present research has made use of formal probability theory and statistics in order to solve the problems of estimation [9]. An schematic example of how a

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

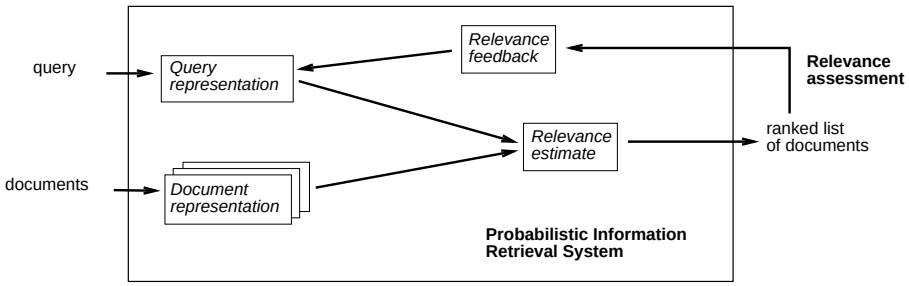

Figure 2: Schematic view of a Probabilistic IR System.

probabilistic IR system works is depicted in figure 2.

Currently there is no PIRS with speech input/output capabilities. The adding of these capabilities to a PIRS is not simply a matter of fitting together a few components, since there needs to be some form of feedback between PIRS and VMD so that the speech input received from the user and translated into text can be effectively recognised by the PIRS and used. If, for example, a word in a spoken query is not recognised by the PIRS, but is very close (phonetically) to some other word the PIRS knows about and that is used in a similar context (information that the PIRS can get from statistical analysis of the word occurrence), then the PIRS could suggest to the user to use this new word in the query instead on the unknown one.

Additionally, it is generally recognised in IR that improvement in effectiveness can be achieved by enhancing the interaction between system and user. One of the most advanced techniques used for this purpose is *relevance feedback*. Relevance feedback is a technique that allows a user to interactively express his information requirement by adapting his original query formulation with further information [14]. This additional information is often provided by indicating some relevant documents among the documents retrieved by the system. When a document is marked as relevant the relevance feedback process analyses the text of the document, picking out words that are statistically significant, and modifies the original query accordingly. Relevance feedback is a good technique for specifying an information need, because it releases the user from the burden of having to think up of words for the query and because it limits the effects of errors in the recognition of query words. Instead the user deals with the ideas and concepts contained in the documents. It also fits in well with the known human trait of "I don't know what I want, but I'll know it when I see it". Relevance feedback can also be used to detect words that were wrongly recognised by the VMD. In fact if, for example, a query word uttered by the user never appears in documents that the user points out to be relevant, while another word similarly spelled (or pronounced) often occurs, then it is quite likely (and we can evaluate this probability) that the VMD was wrong in recognising that word and that

the word really uttered by the user is the other word. These ideas are currently being implemented and tested. The overall performance of the interactive system can be enhanced also using other similar forms of interactions between PIRS and VMD that are current object of study.

3.3 The Document Summarisation System

The *DSS* performs a query oriented document summarisation aimed at stressing the relation between the document content and the query. This enables the user to assess in a better and faster way the relevance of the document to his information need and provide correct feedback to the PIRS on the relevance of the document. Query oriented document summarisation attempts to concentrate users' attention on the parts of the text that possess a high density of relevant information. This emerging area of research has its origins in methods known as passage retrieval (see for example [4]). These methods identify and present to the user individual text passages that are more focused toward particular information needs than the full document texts. The main advantage of these approaches is that they provide an intuitive overview of the distribution of the relevant pieces of information within the documents. As a result, it is easier for users to decide on the relevance of the retrieved documents to their queries.

The summarisation system employed in IVIRS is the one developed by Tombros and Sanderson at Glasgow University [30]. The system is based on a number of sentence extraction methods that utilise information both from the documents of the collection and from the queries used. A thorough description of the system can be found in [30] and will not be reported here. In summary, a document passes through a parsing system, and as a result a score for each sentence of the document is 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 (to distinguish word in the title or in other important parts of the document, for example), and by utilising within-document word frequency information. The document summary is then generated by selecting the top-scoring sentences, and outputting them in the order in which they appear in the full document. The summary length is a fraction of the full document's length. For the documents used in our current implementation of IVIRS the summary length was set at 15% of the document's length, up to a maximum of five sentences.

3.4 The Document Delivery System

The *DDS* performs the delivery of all or parts of the document(s) requested by the user. The user can decide the way and format of the document delivery; this information is usually stored in a user profile, so that the user does not need to give this information every time he uses the system. Documents can be delivered by voice (read in their entirety to the user through the telephone), electronic mail, postal service, or fax; if delivered by electronic mail a number of different document formats are available, like for example, PDF, postscript, or ASCII.

4 Prototype Implementation

The implementation of the prototype system outlined in the previous sections required, 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. A second step involved the development of a model for the VDM and of its interaction with the other components. 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.), and is currently still in progress.

In the implementation of IVIRS we followed a "divide et impera" approach, consisting of dividing the implementation and experimentation of IVIRS in the parallel implementation and experimentation of its different components. The integration of the various components is the last stage.

5 Assessing the Limitations of Current Speech Technology

From the early stages of the implementation of IVIRS, it was clear that current speech technology had a number of limitations that would affect the effectiveness of the system. Although we used state of the art speech recognition and synthesis systems, their use in such a complex system as IVIRS would need to be tested for feasibility.

In this section we report the findings of our feasibility study aimed at addressing the effectiveness of some of IVIRS components in the face of the limitations of

current speech technology. We believe our findings to be of interest and useful to designers and implementors of interactive vocal information access systems.

5.1 User's Perception of Relevance of Spoken Documents

One of the underlying assumptions of the design and development of the IVIRS system is that a user would be able to assess the relevance of a retrieved document by hearing a synthesised voice reading a brief description of its semantic content through a telephone. This is essential for an effective use of the system. In order to test the validity of this assumption we carried out a series of experiments with the DSS and Text-to-Speech module of the IVIRS. The aim of this experimentation was to investigate the effect that different forms of presentation of document descriptions have on users' perception of the relevance of a document. In particular, we were interested in assessing the effectiveness of the use of speech synthesis for the presentation to the user of the retrieved document descriptions.

In a previous study, Tombros and Sanderson [30] used document titles, and automatically generated, query biased summaries as document descriptions, and measured user performance in relevance judgements when the descriptions were displayed on a computer screen and read by the users. The results from that study were used in this experiment, and compared to results obtained when users are listening to the document descriptions instead of reading them. Three different ways of auditory transmission are employed in our study: document descriptions are read by a human to the subjects, read by a human to the subjects over the telephone, and finally read by a speech synthesis system over the telephone to the subjects. The objective was that by manipulating the level of the independent variable of the experiment (the form of presentation), we could examine the value of the dependent variable of the experiment (the user performance in relevance judgements). We also tried to prove that any variation in user performance between the experimental conditions was to be attributed only to the change of level of the independent variable. In order to be able to make such a claim, we had to ensure that the so-called "situational variables" (e.g. background noise, equipment used, experimenter's behaviour) were held constant throughout the experimental procedure. Such variables could introduce bias in the results if they systematically changed from one experimental condition to another [19].

In order to be able to use the experimental results reported in [30], the same task was introduced in our design: users were presented with a list documents retrieved in response to a query, and had to identify as many documents relevant to that particular query as possible within 5 minutes. The information that was presented for each document was its title, and its automatically generated,

query oriented description. Moreover, we used exactly the same set of queries (50 randomly chosen TREC queries), set of retrieved documents for each query (the 50 top-ranked documents), and document descriptions as in [30]. 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. A group consisting of 10 users was employed in the experimental procedure. The population was drawn from postgraduate students in computer science. All users performed the same retrieval task described in the previous paragraph under the three different experimental conditions.

The experiment involved the presentation of document descriptions to subjects in three different forms, all of which were of auditory nature. In two of the experimental conditions the same human read the document descriptions to each subject, either by being physically in the same room (though not directly facing the subject), or by being located in a different room and reading the descriptions over the telephone. In the last experimental condition a speech synthesiser read the document description to the user over the telephone. The speaking rate was kept as constant as possible throughout the whole experiment. 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 it would proceed to the next description.

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 (R) 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 (P) 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 (T), in seconds, that a user takes to assess the relevance of a single document.

A word of caution is perhaps needed here regarding the concept of relevance used in our evaluation methodology. 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;

	S	V	T	C
Avg. P.	47.15	41.33	43.94	42.27
Avg. R.	64.84	60.31	52.61	49.62
Avg. T.	17.64	21.55	21.69	25.48

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

- *user-utility*, which deals with the ultimate usefulness of the piece of information to the user who submitted the query.

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. 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. In the work reported here we were 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. According to this perspective, 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 in completely different ways. Nevertheless, there is seems to be no other way of evaluating auditory interfaces [17].

Table 1 reports the results of the user relevance assessments for all four conditions: on screen description (S), read description (V), description read over the telephone (T), and computer synthesised description read over the telephone (C). These results show how users in condition S performed better in terms of precision and recall and were also faster. Recall and average time slowly decreased from S to C, although some of these differences are not statistically significant (i.e. the average time of V and T). We were surprised to notice that precision was slightly higher for condition T than for conditions V or C. Users tended to be more concentrated when hearing the descriptions read over the telephone than by the same person in the same room, in front of them, but this concentration was not enough when the quality of the voice was getting worse. Nevertheless, the difference in precision between conditions S and C was not so large (only about 5%) to create unsolvable problems for a telephone based IR system.

A difference that was significant was in the average time taken to assess the

relevance for one document. The difference between the condition S and C was quite large and enabled a user to assess on average, in the same amount of time (5 minutes), 70% more documents in condition S than in condition C (22 documents instead of 13). A sensible 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 looking at the documents on the screen, than accessing the system using a telephone. Nonetheless, difference in the average time taken to assess the relevance for one document were very subjective, in fact, we could notice that some users were slow whatever the condition, while other were always fast.

These preliminary results, briefly summarised here, but reported in full in [29], enable us to conclude that presenting document descriptions using speech is indeed feasible and sufficiently effective, even using synthesised speech over a telephone.

5.2 Effects of Word Recognition Errors in Spoken Queries

The effects of word recognition errors (WRE) in spoken documents on the performance of IR systems have been well studied and well documented in recent IR literature. A large part of the research in this direction has been promoted by the SDR track of TREC (see for example [13]). Experimental evidence has brought to the conclusion that for long documents and for reasonable levels of average Word Error Rates (WER), the presence of errors in document transcripts does not constitute a serious problem. In a long document, where terms important to the characterisation of the document content are often repeated several times, the probability that a term will always be misrecognised is quite low and there is a good chance that a term will be correctly recognised at least once. Variations of classical IR weighting schemes (for example giving a lesser importance to the within-document term frequency) have been proposed that are able to cope with reasonable levels of WER [26], but these solutions were found not so effective for short documents.

In contrast with SDR, very little work has been carried out in the area of SQP. SQP is concerned with the use of spoken queries to retrieve textual or spoken documents. To date, very little research work has been devoted to studying the effects of WRE in SQP. A spoken query can be considered similar to a very short document and high levels of WER may have devastating effects on the performance of an IR system. In a query, as in a short document, the misrecognition of a term may cause it to disappear completely from the query representation and, as a consequence, a large set of potentially relevant documents indexed using that term will not be retrieved.

One of the underlying assumptions of the design of IVIRS is that spoken queries could be recognised by the VDM with a level of correctness as to enable their

Data set:	WSJ 1990-92
Number of documents	74.520
Size of collection in Mb	247
Unique terms in documents	123.852
Average document length	550
Average document length (unique terms)	180

Table 2: Characteristics of the Wall Street Journal 1990-92 document collection.

Data set:	Topics 101-135
Number of queries	35
Unique terms in queries	3.304
Average query length (with stopterms)	58
Average query length (without stopterms)	35
Median query length (without stopterms)	28
Average number of relevant documents per query	30

Table 3: Characteristics of topics 101-135 of TREC.

effective use by the PIRS. In order to verify this assumption, we carried out an experimental study of the effects of WRE in spoken queries on the effectiveness of a PIRS. This experimentation is reported in detail in [6, 8]. Here we will only report the most important findings.

In order to experiment the effects of WRE in SQP a suitable test environment needs to be devised. Classical IR evaluation methodology suggests that we use the following: a) a collection of textual document; b) a set of spoken queries with associated relevance assessments recognised at different levels of WER; c) an IR system.

The collection we used is a subset of the collection generated for TREC (see for example [33]). The collection is made of the full text of articles of the Wall Street Journal (years 1990-92). Some of the characteristics of this test collection are reported in table 2.

A set of 35 queries (topics 101-135 of TREC) with the corresponding lists of relevant documents was used. These queries were originally in textual form and were quite long, however some of the fields of the query were not used in the experiments reported in this paper. In fact, the only field used were title, description, and concepts, but considering the text in them indistinctly. This makes the queries short enough to be a somewhat more realistic examples of "real" user queries. Some of characteristics of the queries are reported in table 3.

Query sets:	27	28	29	34	35	47	51	75
Avg. Subst. %	18.8	19.1	20.0	22.7	24.2	31.5	35.5	49.8
Avg. Delet. %	2.6	2.6	2.6	2.6	2.6	3.0	4.2	2.9
Avg. Insert. %	6.0	6.0	6.6	8.4	7.8	12.4	11.3	21.8
Avg. Err. %	27.4	27.7	29.2	33.6	34.6	46.8	51.0	74.5
Avg. Sentence Err. %	39.1	40.0	40.4	42.2	47.0	51.3	56.7	66.5

Table 4: Characteristics of the different query sets.

Since the original queries were in textual form, it was necessary to reproduce then in spoken form and have then recognised by a Speech Recognition (SR) system. This work was carried by Jim Barnett and Stephen Anderson of Dragon Systems Inc. Barnett and Anderson had one single (male) speaker dictate the queries. The spoken queries were then recognised by Dragon's research LVCSR system, a SR system that has a 20,000 vocabulary and a bigram language model trained on the Wall Street Journal data. The bigram language model grows with the vocabulary, and even a reduced vocabulary of 20,000 words can lead to large search spaces represented by the word lattice. Nevertheless the uneven distribution of probabilities among different paths can help. A complexity reduction technique called beam search, consisting of neglecting states whose accumulated score is lower than the best one minus a given threshold, if often used to limit the search. In this way, computation needed to expand bad nodes is avoided. By altering the width of the beam search, a number of sets of transcripts at different levels of WER were generated. The beam width was chosen as the major parameter to alter because it was believed that this yields relatively realistic recognition errors. The standard error characteristics of these sets of transcripts are reported in table 4. We shall refer to these sets as query sets and each set has been denoted by its average WER (i.e. the "Avg. Err. %"). The set identified by 0 (not reported in the table, but reported later on as a reference line) is the perfect transcript. More details on the process used to generate these different sets of transcripts are reported in [2]. This experimental setup, not dependent on telephone applications, was chosen to enable comparison of results with other researchers using this same set of queries. In the future we plan to generate a new set of queries using telephone speech.

Finally, the PIRS used in our study is an experimental IR toolkit developed at Glasgow University which implements a model based on the classical tf_idf weighting schema [24].

In order to analyse the effects of WRE on the effectiveness of SQP, a large number of experiments using the reference PIRS were carried out. In these experiments some of the parameters of the IR process were changed to study their effects on

Query sets:	0	27	28	29	34	35	47	51	75
Avg. P. all	0.22	0.16	0.16	0.16	0.13	0.13	0.14	0.11	0.07
Avg. P. s.q.	0.19	0.13	0.13	0.12	0.10	0.10	0.09	0.05	0.03
Avg. P. l.q.	0.24	0.19	0.20	0.20	0.15	0.16	0.18	0.16	0.11

Table 5: Average precision values for the different query sets, overall, and using long or short queries.

the effectiveness on the SQP task in relation to the different levels of WER. Effectiveness was measured using the standard IR definition of recall and precision. Recall is defined as the portion of all the relevant documents in the collection that has been retrieved. Precision is the portion of retrieved documents that is relevant to the query. However, here we report only the 11-point average precision figures, being these well recognised single-value effectiveness measures. Recall and precision tables are reported in [6, 8].

The first analysis was directed toward studying the effects of different WERs in spoken queries on the effectiveness of a PIRS using a standard text-based parameters configuration. The parameters configuration most commonly used in textual IR employs the tf_idf weighting scheme on terms extracted from documents and queries. Extracted terms are first compared with a stoplist, i.e. a list of non content-bearing terms; terms appearing in the stoplist are removed. The remaining terms are subject to a stemming and conflation process, in order to further reduce the dimensionality of the term space and to avoid a high incidence of the term mismatch problem. In the experiments reported here a standard stoplist [12] and the stemming and conflation algorithm commonly known as "Porter algorithm" were used [23]. Table 5 measures the effects of different WERs in queries on the effectiveness of the PIRS using the above standard configuration.

Naturally, it can be noted that the best results are obtained for the perfect transcript (the transcript 0), and there is a degrade in effectiveness that is related to the WER. Higher WERs cause lower effectiveness. An attentive reader can notice that the reference effectiveness (the one obtained with the perfect transcript) is quite low, especially compared with the level of effectiveness of other IR systems using the same collection, whose performance data can be found in the TREC Proceedings, for example. The reason for these results is due to the fact that no precision enhancement technique, like for example the use of noun phrases or pseudo relevance feedback, was employed in the experimentation reported in this paper. This is a deliberate choice, since it was felt that the use of such techniques would not have allowed a "clean" analysis of the effects of WRE on IR effectiveness.

The first row of table 5 reports the best results obtained in the experimentation,

which are obtained with the *idf* weighting scheme without the use of term frequency information (tf), and without stemming. We can notice that for levels of WERs ranging from 27% to 47% there is no significant difference in performance. Significant low levels of effectiveness can be found for WER 51% and 75%, where the number of errors in the query is so large that what is left of the original query is not enough to work on. One of the possible explanations of this fact can be found in the kind of errors that a SR system produces on the query. It is a known fact that SR produces more errors in short words than in long words. Short words are not very useful for IR purposes, since they are mostly non content-bearing words, many of which can be found in the stoplist. So, as long as the WER is relatively low, mostly short functional terms are affected. When the WER is higher, longer words are affected too and since these words are generally very important in IR, we have a considerable degradation in the effectiveness of the IR process.

Other experiments involving the use of different versions of the tf_idf weighting scheme and of different sizes of stoplists did not produce significantly different results from the ones reported here and will not be presented.

Another series of experiments was conducted to test the robustness of the IR process in relation to query length. It is intuitive to think that the same WER would have a much more detrimental effect on short queries than on long ones. The last two rows of table 5 reports the average precision values for short and long queries. Short queries are queries with less than 28 terms, and long queries those with more than 28 terms, where 28 terms is the median length of a query. The average number of terms in a query, after stopterm removal is 35, therefore there are a number of considerably long queries raising the average. We can notice that short queries have a lower average precision for any level of WER, while long queries have a higher average precisions for any level of WER. This proves the intuition that long queries are more robust to WRE than short queries. For this reason, in the design of the VDM for the IVIRS we will have to exploit dialogue techniques that will elicit the longest possible queries from the users. This is consistent with results of other projects (see for example [21]), and there exists already a number of techniques that we might be able to use in this context [27].

It should be noticed that the study reported here suffers from two important limitations:

1. The queries used are too long and not really representative of typical user queries. Although it has been long recognised that query length is mainly dependent upon the application domain and the IR environment, some initial unpublished user studies on spoken queries indicates that spoken queries are usually longer than written queries.

2. The WERs of the queries used in this experimentation were typical of "dictated" spoken queries, since this was the way they were generated. Dictated speech is considerably different from spontaneous speech and easier to recognise [15]. We should expect spontaneous spoken queries to have higher levels of WER and different kinds of errors.

Despite these limitations , which we plan to tackle with further experimentation, we believe we can conclude that IR is quite robust to WRE in spoken queries. Provided WER in query transcripts remains below 30-40% and the VDM is successful in eliciting long queries, we should expect IVIRS to be effective in the SQP and IR tasks.

6 Overcoming the Limitations of Current Speech Technology

We are currently experimenting with the a number of techniques to improve the effectiveness of SQP for information access. Here we will outline the first results obtained and the directions of the work in progress.

6.1 Relevance Feedback

Given the acceptable level of effectiveness of SQP at levels of WER roughly below 40%, we can conclude that it will be quite likely that in the first n retrieved documents (with n dependent on the user's preference and usually less than 10) there will be some relevant ones. In section 5.1 we have concluded that the user is very likely to be able to perceive the relevance of a document presented in the form of a spoken summary. This result, together with the results reported in section 5.2, enables us to conclude that relevance feedback could be a good strategy to improve effectiveness in a SQP task. The user could find at least one relevant document and feed it back to the IR system which will expand the initial query (therefore also recovering some of the problems due to short queries) and enable to find more relevant documents.

In an initial experimentation, reported in detail in [8], we found that relevance feedback enables to improve the IR effectiveness with spoken queries, in particular for short queries. In fact, as shown in Table 6, the use of relevance feedback enables to reduce considerably the difference in effectiveness between long and short queries. It remains to be seen if users will take advantage of the availability of relevance feedback, since studies on textual IR found that users do not like to use it [14]. The VDM could compel the user to make use of it.

Query sets:	0	27	28	29	34	35	47	51	75
Avg. P. all (1RF)	0.31	0.24	0.23	0.22	0.22	0.20	0.19	0.18	0.16
Avg. P. s.q. (1RF)	0.28	0.23	0.22	0.22	0.21	0.19	0.18	0.15	0.12
Avg. P. l.q. (1RF)	0.32	0.25	0.24	0.23	0.23	0.21	0.20	0.19	0.18
Avg. P. all (2RF)	0.35	0.27	0.27	0.26	0.23	0.21	0.19	0.19	0.15
Avg. P. s.q. (2RF)	0.25	0.26	0.26	0.24	0.20	0.19	0.18	0.17	0.14
Avg. P. l.q. (2RF)	0.35	0.28	0.28	0.27	0.25	0.23	0.20	0.20	0.16

Table 6: Average precision values using relevance feedback with 1 relevant document (1RF) and 2 relevant documents (2RF), for the different query sets, overall, and using long or short queries.

6.2 Use of Prosodic Stress for Topic Detection.

In [25] we investigated the relationship between acoustic stress in spoken sentences and information content. On one side, the *average acoustic stress* was measured for each word throughout each utterance. On the other side an IR index (the tf_idf weight) was calculated. The scatter plots of the two measures showed a correlation between higher values of average acoustic stress with increasing values of the IR index of the word in the majority of the analysed utterances. An example of one such scatter plot is reported in figure 3.

A statistically more valid proof of such a relationship was derived from the histogram of the words with high average acoustic stress vs. the IR index. This confirmed that a word with high average acoustic stress has also a high value of the IR index and, if we trust IR indexes, also high informative content.

The study confirmed our hypothesis of a direct relationship between acoustic stress and information content as identified by IR weighting of spontaneous spoken sentences. The next stage of this work will be the integration of prosodic stress and IR weighting evidence into an new IR weighting algorithm for spontaneous speech. This weighting schema will take into account both word acoustic and statistical clues to characterise the document/query informative content. It will be extremely useful in a number of tasks and in particular for SQP, where the short length of queries requires every possible clue to fully capture the user's information need. A somewhat similar research is currently being undertaken by the speech community for topic segmentation in the topic detection and tracking task [28].

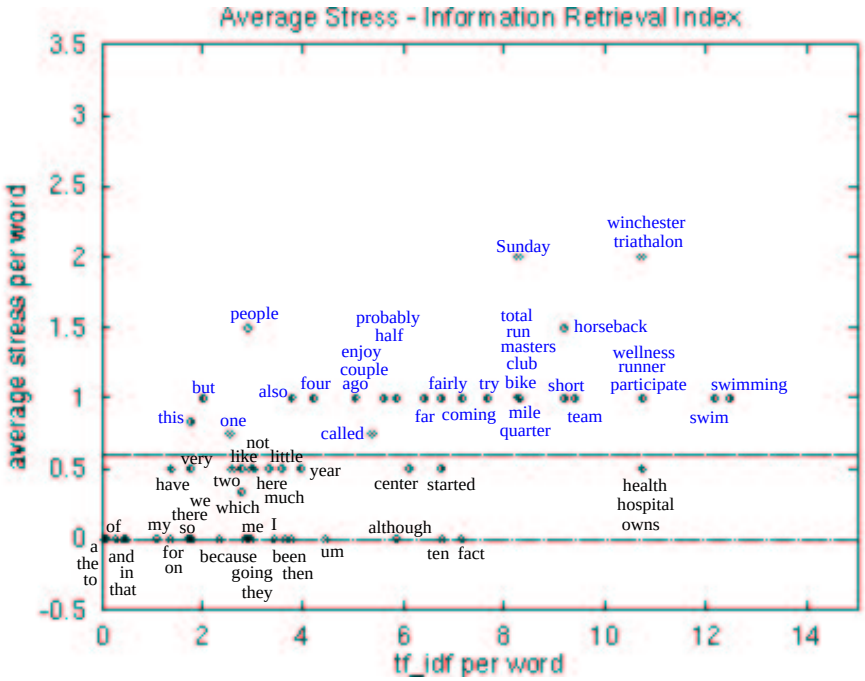

Figure 3: Example of scatter plot of a document were a persons describes how she likes to practice sports and in particular to go swimming on Sundays in Winchester.

6.3 Combination of Semantic and Phonetic Term Similarity

A fundamental problem of IR is "term mismatch". A query is usually a short and incomplete description of the user's information need. Users and authors of documents often use different terms to refer to the same concepts. This fact produces an incorrect relevance ranking of documents with regards to the information need expressed in the query. A similar problem can be found in SDR and SQP, where terms misrecognised by the speech recognition process are found not matching in query and document representations. Naturally, this hinder the effectiveness of the IR system. We called this problem "term misrecognition", by analogy to the term mismatch problem.

In [5] we presented a model for dealing with the term mismatch and the term misrecognition problems in SDR and SQP. Term similarity is used at retrieval time to estimate the relevance of a document in response to a query by looking not only at matching terms, but also at non-matching terms whose semantic and/or *phonetic similarity* are above a predefined threshold. We already proved that semantic similarity, estimated using Expected Mutual Information Measure, can help solve the term mismatch problem [7]. Phonetic similarity, on the other hand,

can help tackle the term misrecognition problem. It can be estimated using Error Recognition Confusion Matrices, for example. An experimental investigation is currently being carried out. The experimental results will provide useful feedback on the effectiveness of the models proposed and on how to effectively combine semantic and phonetic similarity.

7 Conclusions

In this paper we presented the design of an interactive vocal information retrieval system (IVIRS). We also reported some experimental conclusions of a feasibility study related to the use of state of the art speech technology for information access.

The study presented here enables to conclude that, despite apparent limitations, current speech recognition and synthesis technology make it possible to build interactive vocal information access systems that we should expect to be reasonably effective. However, ultimately such effectiveness can only be evaluated using real users carrying out real tasks. As soon as IVIRS will be fully functional we will conduct such an evaluation.

Acknowledgements

Large part of the work reported in this paper was done while the author was at the International Computer Science Institute in Berkeley, California, USA.

References

[1] J. Allan. Perspectives on information retrieval and speech. In A.R. Coden, E.W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer-Verlag, Berlin, Germany, 2002.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo. Experiments in spoken queries for document retrieval. In *Eurospeech 97*, volume 3, pages 1323–1326, Rodhes, Greece, September 1997.

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

[4] J.P. Callan. Passage-level evidence in document retrieval. In *Proceedings of ACM SIGIR*, pages 302–310, Dublin, Ireland, July 1994.

[5] F. Crestani. Combination of semantic and phonetic term similarity for spoken document retrieval and spoken query processing. In *Proceedings of the 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 960–967, Madrid, Spain, July 2000.

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

[7] F. Crestani. Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval*, 2(1):23–43, 2000.

[8] F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Proceedings of Fourth Internation Conference on Flexible Query Answering Systems*, pages 267–281, Warsaw, Poland, October 2000.

[9] F. Crestani, M. Lalmas, C.J. van Rijsbergen, and I. Campbell. Is this document relevant? ...probably. A survey of probabilistic models in Information Retrieval. *ACM Computing Surveys*, 30(4):528–552, 1998.

[10] T. Dutoit. High quality text-to-speech synthesis: an overview. *Journal of Electrical and Electronics Engineering, Australia*, 17(1):25–37, 1997.

[11] T. Dutoit. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[12] W.R. Frakes and R. Baeza-Yates, editors. *Information Retrieval: data structures and algorithms*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

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

[14] D. Harman. Relevance feedback and other query modification techniques. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 11. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[15] D. Jurafsky and J.H. Martin. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 2000.

[16] Y.H. Kao, C.T. Hemphill, B.J. Wheatley, and P.K. Rajasekaran. Toward vocabulary independent telephone speech recognition. In *Proccedings of ICASSP '94*, volume 1, pages 117–120, Adelaide, Australia, April 1994.

[17] J. Kim and W. Oard. The use of speech retrieval systems: a study design. In A.R. Coden, E.W. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, pages 87–93. Springer-Verlag, Berlin, Germany, 2002.

[18] J.A. Markowitz. *Using speech recognition*. Prentice Hall, Upper Saddle River, NJ, USA, 1996.

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

[20] 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.

[21] 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 Grove, CA, USA, February 1991.

[22] J. Peckham. Speech understanding and dialogue over the telephone. In K. Varghese, S. Pfleger, and J-P. Lefevre, editors, *Advanced Speech Applications*, pages 112–125. Springer-Verlag, Berlin, Germany, 1996.

[23] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[24] M. Sanderson. *Word Sense Disambiguation and Information Retrieval*. PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, 1996.

[25] R. Silipo and F. Crestani. Prosodic stress and topic detection in spoken sentences. In *Proceedings of the SPIRE 2000, the Seventh Symposium on String Processing and Information Retrieval*, pages 243–252, La Corunna, Spain, September 2000.

[26] A. Singhal, J. Choi, D. Hindle, D.D. Lewis, and F. Pereira. AT&T at TREC-7. In *Proceedings of the TREC Conference*, pages 239–253, Washington DC, USA, November 1998.

[27] R.W. Smith and D.R. Hipp. *Spoken natural language dialog systems: a practical approach*. Oxford University Press, Oxford, UK, 1994.

[28] A. Stolcke, E. Shriberg, D. Hakkani-Tur, G. Tur, Z. Rivlin, and K. Sonmez. Combining words and speech prosody for automatic topic segmentation. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Washington D.C, USA, 1999.

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

[30] 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.

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

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

[33] E.M. Voorhees and D. Harman. Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the TREC Conference*, pages 1–24, Gaithersburg, MD, USA, November 1998.