Spoken Content Retrieval: A Survey of Techniques and Technologies

Spoken Content Retrieval: A Survey of Techniques and Technologies

Martha Larson

Delft University of Technology, Delft
The Netherlands
m.a.larson@tudelft.nl

Gareth J. F. Jones

Dublin City University, Dublin
Ireland
gjones@computing.dcu.ie

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in
Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is M. Larson and G. J. F. Jones, Spoken Content Retrieval: A Survey of Techniques and Technologies, Foundation and Trends® in Information Retrieval, vol 5, nos 4–5, pp 235–422, 2011

ISBN: 978-1-60198-528-6

Full text available at: http://dx.doi.org/10.1561/1500000020

Foundations and Trends® in Information Retrieval

Volume 5 Issues 4–5, 2011

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2011, Volume 5, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 5, Nos. 4–5 (2011) 235–422
© 2012 M. Larson and G. J. F. Jones
DOI: 10.1561/1500000020

Spoken Content Retrieval: A Survey of Techniques and Technologies

Martha Larson[1] and Gareth J. F. Jones[2]

[1] *Faculty of Electrical Engineering, Mathematics and Computer Science, Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands, m.a.larson@tudelft.nl*

[2] *Centre for Next Generation Localisation, School of Computing, Dublin City University, Dublin, Ireland, gjones@computing.dcu.ie*

Abstract

Speech media, that is, digital audio and video containing spoken content, has blossomed in recent years. Large collections are accruing on the Internet as well as in private and enterprise settings. This growth has motivated extensive research on techniques and technologies that facilitate reliable indexing and retrieval. Spoken content retrieval (SCR) requires the combination of audio and speech processing technologies with methods from information retrieval (IR). SCR research initially investigated planned speech structured in document-like units, but has subsequently shifted focus to more informal spoken content produced spontaneously, outside of the studio and in conversational settings. This survey provides an overview of the field of SCR encompassing component technologies, the relationship of SCR to text IR and automatic speech recognition and user interaction issues. It is

aimed at researchers with backgrounds in speech technology or IR who are seeking deeper insight on how these fields are integrated to support research and development, thus addressing the core challenges of SCR.

Contents

1

Introduction

Spoken Content Retrieval (SCR) provides users with access to digitized audio-visual content with a spoken language component. In recent years, the phenomenon of "speech media," media involving the spoken word, has developed in four important respects.

First, and perhaps most often noted, is the unprecedented volume of stored digital spoken content that has accumulated online and in institutional, enterprise and other private contexts. Speech media collections contain valuable information, but their sheer volume makes this information useless unless spoken audio can be effectively browsed and searched.

Second, the form taken by speech media has grown progressively diverse. Most obviously, speech media includes spoken-word audio collections and collections of video containing spoken content. However, a speech track can accompany an increasingly broad range of media. For example, speech annotation can be associated with images captured with smartphones. Current developments are characterized by dramatic growth in the volume of spoken content that is spontaneous and is recorded outside of the studio, often in conversational settings.

Third, the different functions fulfilled by speech media have increased in variety. The spoken word can be used as a medium for communicating factual information. Examples of this function range from material that has been scripted and produced explicitly as video, such as television documentaries, to material produced for a live audience and then recorded, such as lectures. The spoken word can be used as a historical record. Examples include speech media that records events directly, such as meetings, as well as speech media that captures events that are recounted, such as interviews. The spoken word can also be used as a form of entertainment. The importance of the entertainment function is reflected in creative efforts ranging from professional film to user-generated video on the Internet.

Fourth, user attitudes towards speech media and the use of speech media have evolved greatly. Although privacy concerns dominate, the acceptance of the creation of speech recordings, for example, of call center conversations, has recently grown. Also, users are becoming increasingly acquainted with the concept of the spoken word as a basis on which media can be searched and browsed. The expectation has arisen that access to speech media should be as intuitive, reliable and comfortable as access to conventional text media.

The convergence of these four developments has served to change the playing field. As a result, the present time is one of unprecedented potential for innovative new applications for SCR that will bring benefit to a broad range of users. Search engines and retrieval systems that make use of SCR are better able to connect users with multimedia items that match their needs for information and content.

This survey is motivated by the recognition of the recent growth in the potential of SCR and by our aim to contribute to the realization of that potential. It provides an integrated overview of the techniques and technologies that are available to design and develop state-of-the-art SCR systems. We bring together information from other overviews on the subject of searching speech [5, 25, 38, 76, 91, 148, 166, 180, 193] as well as from a large number of individual research papers. Our survey differs from other overviews in that it encompasses a broad range of application domains and is organized in terms of the overarching challenges that face SCR.

The basic technology used for SCR is Automatic Speech Recognition (ASR), which generates text transcripts from spoken audio. Naïvely formulated, SCR can be considered the application of Information Retrieval (IR) techniques to ASR transcripts. The overarching challenges of SCR present themselves differently in different application domains. This survey takes the position that an SCR system for a particular application domain will be more effective if careful consideration is given to the integration of ASR and IR. The survey provides information on when and how to move beyond a naïve combination of ASR and IR to address the challenges of SCR. Undeniably, ASR has made considerable progress in recent years. However, developing raw technologies and computational power alone will not achieve the aim of making large volumes of speech media content searchable. Rather, it is necessary to understand the nature of the spoken word, spoken word collections and the interplay between ASR and IR technologies, in order to achieve this goal.

Reading the survey. This survey is aimed at the reader with a background in speech technologies or IR who seeks to better understand the challenges of developing algorithms and designing systems that search spoken media. It provides a review of the component technologies and the issues that arise when combining them. Finally, it includes a brief review of user interaction issues, which are key to truly useful SCR systems.

The survey can be read sequentially from beginning to end, but is structured in modules, making it possible to read parts of the survey selectively:

- The present *Introduction* defines SCR as it is used in this survey, and differentiates it from related tasks that fall outside of the survey's scope. Further, it provides a brief overview of SCR research, including a summary of the two-decade history of the field of SCR.
- *Overview of Spoken Content Indexing and Retrieval* begins with the presentation of a general SCR architecture. For completeness, a high-level overview of IR techniques is

provided. Then, in subsection 2.4, “Challenges for SCR,” we set out a list of the key challenges faced in designing and implementing SCR systems. The presentation of SCR techniques and technologies in the remainder of the survey is motivated by the need to address these key challenges.

- *Automatic Speech Recognition* presents, for completeness, a high-level overview of human speech and of ASR technology. Then, issues specific to SCR are addressed in subsection 3.3, “Aspects of ASR Critical for SCR,” and in subsection 3.4, “Considerations for the Combination of ASR and IR.” These subsections focus on specific aspects of ASR and its integration with IR and introduce issues that are covered in greater depth in the rest of the survey.
- *Exploiting Automatic Speech Recognition Output* presents techniques used to exploit ASR within an SCR system, including making use of multiple ASR hypotheses and subword units.
- *Spoken Content Retrieval beyond ASR Transcripts* discusses how ASR output can be supplemented to improve SCR, including issues related to extending ASR transcripts effectively and also to structuring and representing speech media.
- *Accessing Information in Spoken Content* addresses issues involving user interaction with speech media and the presentation of search results to users.
- *Conclusion and Outlook* summarizes the major themes from a high-level perspective and presents an outlook to the future.

A particularly important feature of this survey is its extensive bibliography, including over 300 references. The bibliography was selected with the goal of providing a comprehensive selection of entry points into the literature that would allow further exploration of the issues covered by this survey.

1.1 Definition of Spoken Content Retrieval (SCR)

In the broad sense, SCR encompasses any approach that aims to provide users with access to speech media. However, in the narrow sense, its goal is much more specific. In SCR, "Retrieval" is used as it is in IR, namely, to designate the task of automatically returning content with the aim of satisfying a user information need, expressed by a user query. SCR involves an interpretation of the user need and a matching of that need to the speech media. We formalize this concept with the following definition:

> **Spoken Content Retrieval** is the task of returning speech media results that are relevant to an information need expressed as a user query.

Since the emergence of research related to search of speech media, a number of terms have been used to refer to various tasks and techniques. It is worthwhile highlighting their similarities and differences here. The term "speech retrieval" (SR) was used in the first IR paper to treat SCR [87], which explored search of radio news. This form of SCR soon became generally known as "spoken document retrieval" (SDR). This term is used to refer to retrieval techniques for collections having pre-defined document structure, such as stories in broadcast news. As the field has matured, it has become clear that for many tasks, there is no pre-defined or natural definition of documents and that the term SDR is not always appropriate. The term "speech retrieval" [205] was re-adopted as an umbrella designation for search in collections with and without document boundaries.

At the same time, the field of "voice search" or "voice retrieval" has emerged, which is focused on returning results (which may be textual) to queries that have been spoken by users [285]. In order to clearly distinguish searching speech tasks from spoken-query tasks, the designation "speech-based information retrieval" is used. This designation also serves to emphasize that the results returned to the user may actually have other modalities alongside of spoken content, such as the visual channel in video [202]. Our choice of "Spoken Content Retrieval" encompasses both SDR and SR, while keeping the focus clearly on the spoken word as content, not query, and including not just audio-only speech content, but rather speech media in its wide array of different forms, including video.

1.2 Relationship of SCR to Information Retrieval (IR)

SCR is often characterized as IR performed over text transcripts generated by an ASR system. This survey takes the position that this characterization is too naïve to be useful in every situation. The extent to which it is possible to create an SCR system by indexing the output of an out-of-the-box ASR system using an out-of-the-box IR system will ultimately depend on the domain of application and the use case, including the user tasks, the complexity and content of the data, the types of queries that users issue to the system and the form of results that they expect to receive in return. In this subsection, we discuss SCR issues from the IR perspective.

1.2.1 Differences between SCR and IR

Generally speaking, there are several differences between SCR and text IR that vary to differing degrees depending on the situation. The most often cited difference between SCR and text IR is the fact that transcriptions generated by ASR systems generally contain errors. This can mean that an SCR system will often need to make a collection searchable using ASR transcripts that have a high average error rate, sometimes as high as 50%. Under such conditions, SCR cannot be treated as merely a text IR task since this level of noise in the “text” will impact on IR effectiveness.

An additional difference is that spoken audio, unlike text, is rarely structured into logical units such as paragraphs, or even sentences, meaning that some form of segmentation into retrieval units is often required prior to entering the data into the retrieval system. Also, speech is a temporal medium, meaning that a speech signal extends over a fixed length of time. As a result, accessing raw spoken content is time consuming and inefficient, meaning that SCR systems must provide visualizations of spoken content in results lists and in playback interfaces. Such visualizations allow users to scan and access spoken material efficiently, faster than in real time.

Further, it is important not to overlook the fact that ASR technology can generate information that is not included in standard text media. This information can be exploited by the SCR system and generally comes in several forms. First, each recognized word is accompanied by a time code indicating its position within the speech media. Second, the ASR system generates acoustic information reflecting the closeness of the match between a given word and the speech signal at a particular position. Third, the ASR system generates information about words that were potentially spoken within the speech signal, but were not found by the system to be in the most likely transcription of the signal (so-called multiple hypotheses). Also, when combined with additional audio analysis technology, an ASR system is able to generate rich transcripts that contain more information than text. For example, encoding speaker characteristics such as speaker change points, male/female speaker and identifying the speaker or audio events, such as applause and laughter. We return to issues relevant to the difference between SCR and IR in *Exploiting Automatic Speech Recognition Output* and *Spoken Content Retrieval beyond ASR Transcripts.*

1.2.2 User Information Needs for SCR

An information need can be defined as the reason for which the user turns to a search engine [57]. In our case, the information need is the reason why the user turns to an SCR system. The information need can be thought of as the set of characteristics that an item must possess in order for it to satisfy the requirement that motivated the user to engage in a search activity. In general, the sorts of characteristics desired by users determine the approaches that are best deployed by the SCR system. Assumptions about the nature of user needs inform the design process of an SCR system. The more explicit these assumptions can be made, the more likely the SCR system will succeed in fulfilling user needs. For example, if it is safe to assume that users will be satisfied with segments of audio in which a speaker has pronounced the query term or terms, then the SCR system should be implemented as a system that detects the location of mentions of specific spoken terms. From this most basic “finding mention” type of speech search, systems should grow more complex, only to the extent that it is necessary in order to meet the user needs.

In [298], it is noted that speech retrieval systems have conventionally paid little attention to user requirements. Here, we mention a handful of examples of research papers on systems, which give a clear statement of the nature of the user need that the systems are designed to handle. An early example is the voice message routing application in [231], which specifies that the system is intended to sort voice messages or route incoming customer telephone calls to customer service areas. In [94], the design of an SCR system for a large oral history archive is described. User requirement studies were performed that made use of actual requests that had been submitted to the archives and also of the literature concerning how historians work with oral history transcripts. In [21], a user study is conducted for the domain of podcasts, and five different user goals in podcast search are identified and used as the basis for evaluation of an SCR system.

The reasons that motivate users to turn to speech search are diverse. It is arguable that the range of user search goals for SCR is larger than for traditional text-based IR. Consider the example query, `taxes lipreading`. Two possible information needs behind this query are: “Find results discussing George Bush’s famous quote, *Read my lips, no new taxes*” and “Find items discussing recent decisions by the Federal Communications Commission to impose a fee on Video Relay Service for the deaf.” It is clear that the query either under-specifies or misspecifies the information need and that the IR system will have a serious burden of query interpretation. However, the possibilities are multiplied if the collection to be searched contains speech media rather than text. In addition to these two information needs, the following could also be possible, “Find items in which a speaker pronounces the phrase *Read my lips, no new taxes*” and “Find a recording of the original speech in which Bush said *Read my lips, no new taxes*.”

In order to satisfy user information needs, an SCR system must also fulfill user interaction requirements. In general, it is not sufficient that the SCR system returns items that are good matches to the user information need. Rather, the system must also present an item in a way that also convinces users that it is a good match. Users do not examine all results in detail, and are very likely to skip over results that, at the first glance, look like they will not be useful. The effect is particularly egregious in the case of SCR, due to the time that it takes to "listen-in" to particular spoken content hits or view individual segments of video. We will return to these issues in more detail in *Spoken Content Retrieval beyond ASR Transcripts* and *Accessing Information in Spoken Content.*

1.3 Relationship of SCR to Speech Recognition

In this subsection, we discuss SCR issues from the ASR perspective. Speech recognition research naturally falls into two main branches. The first branch, called Speech Understanding (SU), is devoted to developing dialogue systems capable of carrying on conversations with humans for the purpose of, for example, providing train schedule information. The second branch is arguably the more closely related to SCR and has performed research in the "listening typewriter" speech transcription paradigm. Under this paradigm, given a stream of speech, the goal of the ASR system is to generate a transcript of the words spoken, equivalent to one that would be made by a human sitting at a typewriter. In this paradigm, the ASR system should operate as independently as possible from the domain or the topic of speech. By contrast, SU systems typically operate in highly constrained domains and involve complex models intended to capture and exploit the semantic intent of the speaker.

Recently, the field of ASR has been moving away from the "listening typewriter" paradigm and towards forms of speech output that are specifically designed to provide indexing terms (words and phrases) that can be used as the basis of SCR. Early systems used a fixed set of keywords and identified spoken instances of these keywords in the speech stream, a task referred to as "wordspotting" [301]. Further development in this area was devoted to dropping the restriction that the keywords must be specified in advance [122]. More recently, the keyword spotting paradigm has attracted renewed interest dedicated to creating efficient systems capable of handling large amounts of spoken content. For such systems, the designation "Spoken Term Detection" (STD) is generally applied [199]. The STD task returns instances of particular words being pronounced within the speech stream. A related task, Spoken Utterance Retrieval (SUR), involves returning short documents in which specific words are pronounced. If STD or SUR is used to search for particular query words, it can be considered a form of speech search, or even retrieval. However, STD and SUR are, in and of themselves, blind to larger meaning. In other words, systems designed to carry out these tasks make no attempt to match results with an underlying need for a specific sort of content (e.g., content on a particular topic) expressed by the user query. In order to match speech media and user needs, SCR is necessary. In the next subsection, we develop a systematic comparison between tasks closely related to ASR and those that are from core SCR tasks.

1.4 SCR and Other "Searching Speech" Tasks

It is possible to identify a large range of "searching speech" tasks that are similar to SCR in that they can be characterized by the same surface form (i.e., matching a string to speech content) and also make use of the same underlying technology (i.e., ASR). These tasks are related to SCR, but are distinct from the core case of SCR that is the topic of this survey. We distinguish between four different tasks, summarized in Table 1.1, that have the same surface form as SCR and make use of ASR technology.

The four tasks are broken down along two dimensions. The first dimension involves how the system addresses the user need, that is, the criteria by which the match between the user query and the spoken content items is determined. In a "finding mentions" type task, the

Table 1.1. ASR-based search takes the form of four tasks, involving two dimensions.

	User need known at **indexing time**	User need known at **search time**
System addresses need by finding **mentions** (words or phrases)	wordspotting	spoken term detection (STD)
System addresses need by finding **relevant content** (documents, segments, entry points)	classification filtering	spoken content retrieval (SCR)

user inputs a query and the system returns occurrences of the query string found within the ASR transcript. This type of task includes wordspotting, STD and SUR. In a "finding mentions" task, a hit is considered to be a successful match to the query if it contains the words or the query string pronounced in the speech stream. A mention can be returned as a result to the user in the form of either a timepoint (i.e., for STD) or a larger item containing that string (i.e., for SUR). In a "finding content" task, the user inputs a query and the system returns items that either treat the topic specified by that query or fit the description of that query. We consider the core case of SCR to be "finding content" tasks. The importance of the "finding content" SCR task is also emphasized in [38], which refers to it as "evaluating performance from a document retrieval point of view" (p. 42).

It is important to recognize that for a "finding mentions" task and for a "finding content" task the input string (i.e., the query) can be identical. The difference lies in how the retrieval system interprets the information need behind this query. A simple example illustrates the difference. Under an SCR scenario, the retrieval system would respond to the query `volcanic ash`, by providing results that explain the properties, causes and effects of volcanic ash. If a speaker utters the sentence, "The organizers put together a diverse and interesting program and the Future Internet Assembly was a great success, despite air travel interruption due to volcanic ash," the appearance of the phrase "volcanic ash" in that utterance would not necessarily be sufficient to constitute relevance for an SCR result. The topic of this utterance is the Future Internet Assembly, and it is likely to be more directly relevant to queries concerning this event. Under an STD scenario, however, this phrase would clearly be relevant to the query `volcanic ash` since it contains the spoken phrase "volcanic ash." If the system failed to return this occurrence as a result, the STD system would be considered to have failed to retrieve a relevant result.

The second dimension involves prior availability of the information concerning the requests to which the system is expected to provide a response. The first category under this dimension comprises tasks that have information about the user need at indexing time, that is, at the moment at which indexing features for the spoken content items are generated. Early wordspotting systems are "finding mention" systems, which fall into this category. Here, the information need is fixed in the form of a list of terms that must be found in the spoken content stream. As noted earlier, for such wordspotting systems, this list must be known at "ASR-time," that is, the moment at which the ASR transcripts are produced. Spoken content classification and filtering systems also fall into this category. Here, the information need is constituted by a topic class and the system is provided in advance with a list of classes it is expected to identify. The classification system judges the content of the speech media items and makes a decision on whether or not each item belongs to a class. Note that spoken content classification is a "finding content" type task, thus mention of the name of the class (e.g., "cooking") in the speech media item is not enough to guarantee membership in that class. The item must actually treat subject material that belongs to that topical class. Typically, labeled training data are used to train classifiers that are able to separate in-class from out-of-class items.

The second category under this dimension comprises tasks for which no information about the user need or query is available until search time. Early wordspotting systems quickly evolved into keyword spotting systems that required no advance knowledge of the query. Currently, keyword spotting techniques are researched in the context of either STD or SUR. The core case of SCR is a "finding content" task in which there is no information available in advance. In sum, although SCR is clearly related to other "searching speech" tasks, it is distinct in that it involves responding to the information need, i.e., the topical specification or the item description, represented by an ad hoc query posed by the user.

"Searching speech" tasks also differ according to whether the spoken content collection is treated as static, or relatively static, or whether it involves a steady stream of incoming spoken content. Thus far, we have discussed tasks that involve a static collection, one that does not grow over time. In another scenario, the collection is dynamic, that is, new speech content is constantly arriving and the goal of the system is to make a judgment about the incoming stream. Such a task is referred to as *information filtering* or *media monitoring*. The information need can consist of finding mention, or it can consist of identifying topics.

If new topics must be discovered within the stream, the task is often referred to as Topic Detection and Tracking (TDT) [4].

It is important to note that although the tasks in Table 1.1 cannot all be considered core cases of SCR, they all make an important contribution to SCR. As has been noted, these tasks are all "searching speech" tasks, that is, they are related via their surface form and their use of ASR. However, there is a further connection that motivates us to include discussion of these tasks in this survey: these tasks can be used as sub-components of an SCR system whose function it is to extract indexing features that will be used for the purposes of retrieval. We will return to mention these tasks again in *Exploiting Automatic Speech Recognition Output* and *Spoken Content Retrieval beyond ASR Transcripts.*

1.4.1 Other Tasks Related to SCR

We now proceed to briefly treat two other tasks that are often mentioned in the context of searching speech, but which do not fall into the scope of this survey.

Spoken queries/Query by example. Spoken queries can be used to query either a text collection or a speech media collection. In either case, if the query is short, a word error in the query can be difficult to compensate for. Systems that accept spoken queries are often referred to as "voice search" systems. Work on spoken queries includes [15, 151, 285]. Research comparing spoken to written queries is described in [56, 191]. Finally, a technique that bears affinity with spoken query techniques is query by example [188, 262]. Here the information need of the user is specified with a sample of the types of documents that are relevant and the system returns documents that match these samples on the basis of spoken content. Techniques in which the user need is expressed as speech are clearly relevant for SCR, but will not be treated as part of the material covered in this survey.

Question answering. The task of question answering (QA) involves extracting the answer to a user's question from an information source. There has been very extensive work on QA from text sources in recent years. However, there is also interest in developing QA for spoken data. For example QA for lectures and meetings has been reported in [55, 233], while [310] describes research on video news QA. QA for spoken data can utilize many of the methods developed for text QA. However, as with the application of any natural language processing techniques to speech, the noise in ASR transcripts must be taken into account. This may require methods to be simplified for application to speech data. In terms of answer presentation, this could simply make use of the ASR transcript. Alternatively a portion of the audio could be played back. In the case of the latter option, the potential need to provide context to enable the user to understand what is being said must be taken into account. Question answering is also quite evidently related to SCR.

1.5 A Brief Overview of SCR Research

Research in SCR and its underlying technologies has been ongoing for more than twenty years. During this time many techniques have been proposed and explored for different tasks and datasets. This subsection begins with a brief chronological history of SCR research from its birth to the present. We then offer an overview of some application areas and a brief discussion of SCR research for different languages of the world. Our objective here is both to present a historical perspective of the development of SCR and to highlight the key technological innovations at each point.

1.5.1 The History of SCR Research

The history of SCR research falls relatively neatly into four different eras. Each new era brought new tasks, new algorithms and new initiatives to strengthen the SCR research community.

The first era can be thought of as *Proto-SCR* and its heyday was in the early 1990s. Key examples of work conducted in this era are [230, 231] from MIT Lincoln Labs and [301] from Xerox PARC. Modern large-vocabulary continuous speech recognition (LVCSR) had not yet emerged onto the scene, and systems addressed the task of filtering voice messages by using wordspotting techniques, which recognized a small set of words within the speech stream. In [230, 231], the task is referred to in the literature as "information retrieval," but it differs from the concept of IR as understood by the IR research community. In this work, topics or speech message classes were defined ahead of time and not at the time at which the system was queried. Instead, this task is more akin to "information filtering" (cf. subsection 1.4) than SCR.

We call the second era the *Dawn of SCR*. This era arguably began with the 1992 publication of [87], a description of a prototype "System for Retrieving Speech Documents" at ETH Zürich. The prototype made use of subword indexing features and, critically, information about the queries or the information needs of the users did not have to be available to the system in advance. Other systems dating from this era also accepted ad hoc queries from users. The year 1994 saw the publication of [122], which proposed a wordspotting approach based on phonetic lattices that made it possible to carry out vocabulary-independent wordspotting after recognition. If an LVCSR system alone is used to transcribe the spoken content, the index of the SCR system will be limited to containing those words occurring in the vocabulary of the recognizer. Phone Lattice Spotting (PLS) made possible vocabulary independent SCR and was exploited by subsequent work at Cambridge University [27, 120, 121]. An important result to emerge was that the vocabulary independence of PLS could be combined with the robustness of LVCSR to obtain improved SCR results [132]. This era was characterized by research conducted in isolation at individual research sites. During this era, the first systems for broadcast news retrieval were an important development, especially the Informedia system at Carnegie Mellon University (CMU) [105]. The Informedia project established the first large scale digital video search system, with its search driven by a combination of manually-generated closed captions and LVCSR transcriptions.

The SCR research scene changed dramatically with the beginning of what we refer to as the *Rise of the SCR benchmark*. This era dates from 1997, the year the Text REtrieval Conference (TREC) [277] offered the first SDR task. Research sites emerged from isolation as they began working on the same data sets and tasks within the framework of benchmark initiatives. This focus made it possible to compare results across algorithms and across sites. The TREC tasks provoked a variety of research into methods to improve SCR effectiveness, notably the value of query expansion [306] for SCR and an exploration of document expansion [251]. This era drew to a close with the publication in 2000 of [82], which broadly concluded that the problems of SCR, as defined in terms of retrieval of spoken documents, were either solved or sufficiently well characterized to be addressed without significant further research effort. Remaining challenges were identified as involving more complex tasks, such as question answering or spoken queries, or extending the environment to multimedia video search.

The present era can be characterized as the era of *Spontaneous, conversational speech*. It can be considered to have begun in 2001, with a workshop entitled "Information Retrieval Techniques for Speech Applications" [53] organized at the ACM SIGIR (Special Interest Group on Information Retrieval) Conference, at which the keynote speaker [3] pointed out that TREC SDR had focused on long documents and long queries, in contrast to the shorter queries or shorter documents characterizing many of the new SCR use scenarios. In such scenarios, the importance of speech recognition error could rise enormously. Arguably, however, the era of spontaneous, conversational speech did not get under way until there was also a spontaneous, conversational benchmark task available to provide researchers with material to experiment and compare results. In 2005, a Spoken Retrieval track organized within the Cross-Language Evaluation Forum (CLEF) used a large, challenging corpus of interview data [205]. In 2008, a video retrieval track was founded within CLEF, which later developed into an independent benchmark called MediaEval. This benchmark offers tasks that make use of user-contributed speech media collected from the Internet [156, 160]. The TREC Video Retrieval Evaulation (TRECVid) benchmark [255] has conventionally focused on the visual relevance of video to user queries, but makes use of ASR transcripts and has recently expanded the notions of relevance that it explores. A further evaluation involving search of informal speech was introduced in 2011 at the 9th NTCIR: NII Testbeds and Community for Information access Research Evaluation Workshop, where the SpokenDoc track had STD and SCR tasks focused on searching a corpus of Japanese lectures [1].

1.5.2 Use Scenarios

SCR has been applied in a range of different application areas. Initially, the dominant application was access to broadcast news data. This research first investigated radio news [87, 121] and later television news in the Informedia project [105]. It formed the basis of the TREC SDR datasets [82]. Broadcast media reports involve a combination of scripted and unscripted material, however, they are well-behaved in the sense that the topical scope is limited and they have an underlying structure that is readily identifiable.

Another area that received considerable attention in the early phases of SCR research was voice mail. Both the SCANMail project [297, 298] and the Video Mail Retrieval using Voice project [27, 132] focused on search of spoken mail messages. Of particular note are the studies at AT&T that explored users' interaction with audio content from a cognitive perspective. These studies investigated, for example, people's poor ability in recalling details in spoken content, such as answering machine messages [112, 113]. Understanding how people actually interact most effectively without audio material is crucial to the success of SCR systems.

Other application areas involve less planned, more spontaneous speech or speech that is produced within the context of a conversation or other less formal settings. Search of this less well-planned content has formed the basis of more recent work in SCR. Examples include search of meetings, [23, 150], call center recordings [176], collections of interviews [33, 61], historical archives [100], lectures [86], podcasts [207], and political speeches [2].

In [38], it is noted that the best method for indexing audio data can differ according to the goal of the retrieval system. For this reason, a good understanding of the underlying use scenario will translate into a more highly effective SCR system. Differences between use scenarios encompass both differences between user needs and differences between the underlying spoken content collection, in terms of language, speaking style, topic, stability and structure.

1.5.3 Languages of the World

With the notable exception of the work at ETH Zürich [87, 88], the early work on SCR was devoted to English language spoken content. Retrieval of English language content is relatively simple, since features to be searched in the form of words are readily available. Some pre-processing in the form of stemming, see *Overview of Spoken Content Indexing and Retrieval*, can be used to match different word forms, but the features themselves are easily identified. This is not the case for many other languages. For example compounding languages (e.g., German and Dutch) express semantically complex concepts using single lexical words. These must often be de-compounded to constitute simpler words for search. Still more challenging are agglutinative languages (e.g., Turkish and Finnish), which have enormous vocabularies resulting from the combination of a relatively small set of morphemes with a vocabulary of stems. Extracting suitable search features for these languages can be a complex process. In the case of languages like Chinese, where whitespace is not used to delimit words in written language, segmentation methods are required. Generally, a separate ASR system must be deployed for every language that is to be included in a spoken content index. An ASR system for a new language involves a large implementation effort and in some cases an optimization of the basic design of the ASR system. These issues are addressed in greater depth in *Exploiting Automatic Speech Recognition Output.*

Although much of the research discussed in this survey has been carried out for English-language spoken content, we would like to emphasize the importance of considering the full scope and variety of human languages for research and development in SCR. Research on SCR for non-English languages is gaining in volume. Coverage for a number of languages has been relatively strong. Work on non-English SCR includes: Chinese [283], German [155, 241], Italian [71], French [84], Dutch [212], Finnish [153], Czech [200], Japanese [167], and Turkish [8].

Cross-language speech retrieval combines SCR with machine translation techniques in order to give users querying in one language access to speech collections in another language. Such a system is helpful for users who have passive knowledge of a language, and would be able to derive benefit from listening to or watching speech media in that language, but whose knowledge is not advanced enough to allow them to formulate queries. As noted by [133], scenarios for cross-language speech retrieval include cases in which the collection contains multiple languages or accepts queries formulated in multiple languages. Early work in the area of cross-language SCR includes [216], which describes a system that accepts a textual query in French and returns spoken German broadcast news stories. Much of the work on cross-language speech retrieval has been carried out within the CLEF [217]. Other important work includes that on Mandarin Chinese/English cross-language speech retrieval [183].

Now we turn to a more detailed overview of spoken content indexing and retrieval, including a high-level overview of IR techniques, which will allow us to formulate a list of the key challenges faced when designing and implementing SCR systems.

References

[1] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for spoken documents task in NTCIR-9 Workshop," in *Proceedings of the NII Test Collection for IR Systems Workshop*, pp. 223–235, 2011.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4873–4876, 2009.

[3] J. Allan, "Perspectives on information retrieval and speech," in *Information Retrieval Techniques for Speech Applications*, (A. R. Coden, E. W. Brown, and S. Srinivasan, eds.), pp. 323–326, Springer Berlin/Heidelberg, 2002.

[4] J. Allan, "Topic detection and tracking: Event-based information organization," in *The Kluwer International Series on Information Retrieval*, vol. 12, Springer, 2002.

[5] J. Allan, "Robust techniques for organizing and retrieving spoken documents," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 1, pp. 103–114, 2003.

[6] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proceedings of the NIST Machine Learning for Multimodal Interaction, Meeting Recognition Workshop*, pp. 26–38, 2005.

[7] J. Archibald and W. O'Grady, *Contemporary Linguistics.* Bedford/St. Martin's, 2001.

[8] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

[9] B. Arons, "SpeechSkimmer: Interactively skimming recorded speech," in *Proceedings of the ACM User Interface Software and Technology Conference*, Atlanta, 1993.

[10] B. Arons, "SpeechSkimmer: A system for interactively skimming recorded speech," *Transactions on Computer Human Interaction*, vol. 4, no. 1, pp. 3–38, 1997.

[11] B. Arons and E. Mynatt, "The future of speech and audio in the interface: A CHI '94 workshop," *SIGCHI Bulletin*, vol. 26, no. 4, pp. 44–48, 1994.

[12] X. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 89–114, 2002.

[13] C. Auzanne, J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Automatic language model adaptation for spoken document retrieval," in *Proceedings of the RIAO Conference on Content-Based Multimedia Information Access*, pp. 132–141, 2000.

[14] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search.* Addison-Wesley Longman Publishing Co., Inc., 2010.

[15] B.-R. Bai, L.-F. Chien, and L.-S. Lee, "Very-large-vocabulary Mandarin voice message file retrieval using speech queries," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1950–1953, 1996.

[16] J. Baker, "The DRAGON system — an overview," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[17] S. Banerjee and A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings," in *Proceedings of Interspeech*, 2006.

[18] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[19] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw, "Combining the evidence of multiple query representations for information retrieval," *Information Processing & Management*, vol. 31, no. 3, pp. 431–448, 1995.

[20] M. Benzeguiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and intrinsic speech variation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. V/1021–V/1024, 2006.

[21] J. Besser, M. Larson, and K. Hofmann, "Podcast search: User goals and retrieval technologies," *Online Information Review*, vol. 34, p. 3, 2010.

[22] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, pp. 205–231, May 1996.

[23] H. Bourlard and S. Renals, "Recognition and understanding of meetings overview of the European AMI and AMIDA projects," IDIAP-RR 27 Technical Report, 2008.

[24] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[25] E. W. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, and A. Amir, "Toward speech as a knowledge resource," *IBM Systems Journal*, vol. 40, no. 4, pp. 985–1001, 2001.

[26] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "Automatic content-based retrieval of broadcast news," in *Proceedings of the Annual ACM International Conference on Multimedia*, pp. 35–43, 1995.

[27] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 307–316, 1996.

[28] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young, "Video mail retrieval using voice: An overview of the Cambridge/Olivetti retrieval system," in *Proceedings of the ACM Multimedia Workshop on Multimedia Database Management Systems*, pp. 47–55, 1994.

[29] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young, "Automatic content-based retrieval of broadcast news," in *Proceedings of the Third ACM International Conference on Multimedia*, pp. 35–43, 1995.

[30] C. Buckley, G. Salton, J. Allan, and A. Singha, "Automatic query expansion using SMART: TREC 3," in *Proceedings of the Third Text Retrieval Conference*, pp. 69–80, 1995.

[31] J. Butzberger, H. Murveit, E. Shriberg, and P. Price, "Spontaneous speech effects in large vocabulary speech recognition applications," in *Proceedings of the Workshop on Speech and Natural Language*, pp. 339–343, 1992.

[32] S. Büuttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[33] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, vol. 12, no. 4, pp. 420–435, 2004.

[34] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, K. Vasilis, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, Chapter 3, pp. 28–39, Springer, 2006.

[35] J. Carmichael, M. Larson, J. Marlow, E. Newman, P. Clough, O. Oomen, and S. Sav, "Multimodal indexing of digital audio-visual documents: A Case study for cultural heritage data," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, pp. 93–100, London, U.K., 2008.

[36] J. K. Chambers, P.Trudgill, and N. Schilling-Estes, eds., *The Handbook of Language Variation and Change*, Blackwell Handbooks in Linguistics. Wiley-Blackwell, 2004.

[37] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pp. 443–450, Morristown, NJ, USA, 2005.

[38] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[39] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech and Language*, vol. 21, no. 3, pp. 458–478, 2007.

[40] B. Chen, "Exploring the use of latent topical information for statistical Chinese spoken document retrieval," *Pattern Recognition Letters*, vol. 27, no. 1, pp. 9–18, 2006.

[41] B. Chen, H.-M. Wang, and L.-S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 303–314, 2002.

[42] F. R. Chen and M. Withgott, "The use of emphasis to automatically summarize a spoken discourse," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/229–I/232, 1992.

[43] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, no. 4, pp. 359–393, 1999.

[44] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[45] Y.-T. Chen, B. Chen, and H.-M. Wang, "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 95–106, 2009.

[46] T. Cheong, R. Kok, J. Schuurman, and B. Stukart, "Improving the front-end of Kunststofzuiger," Final Report Project Information Retrieval, University of Amsterdam, 2008.

[47] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "Statistical lattice-based spoken document retrieval," *ACM Transactions on Information Systems*, vol. 28, no. 1, pp. 1–30, 2010.

[48] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*, pp. 26–33, 2000.

[49] M. G. Christel and R. Yan, "Merging storyboard strategies and automatic retrieval for improving interactive video search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 486–493, 2007.

[50] K. W. Church, "Speech and language processing: Can we use the past to predict the future?," in *Text, Speech and Dialogue*, vol. 3206 of *Lecture Notes in Computer Science*, (P. Sojka, I. Kopecek, and K. Pala, eds.), pp. 3–13, Springer Berlin/Heidelberg, 2004.

[51] J. Clark, C. Yallop, and J. Fletcher, *An Introduction to Phonetics and Phonology (Blackwell Textbooks in Linguistics)*. Wiley-Blackwell, 2007.

[52] A. R. Coden and E. W. Brown, "Speech transcript analysis for automatic search," in *Proceedings of the Annual Hawaii International Conference on System Sciences, 2001*, 2001.

[53] A. R. Coden, E. W. Brown, and S. Srinivasan, "ACM SIGIR 2001 workshop "Information Retrieval Techniques for Speech Applications"," *SIGIR Forum*, vol. 36, no. 1, pp. 10–13, 2002.

[54] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, L. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spiitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue, "The challenge of spoken language systems: Research directions for the nineties," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 1–21, 1995.

[55] P. R. Comas, J. Turmo, and L. Marquez, "Sibyl, a factoid question answering system for spoken documents," *ACM Transactions on Information Systems*, vol. 30, no. 3, 2012.

[56] F. Crestani and H. Du, "Written versus spoken queries: A qualitative and quantitative comparative analysis," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 881–890, 2006.

[57] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, 1st Edition, February 2009.

[58] T. H. Crystal, A. Schmidt-Nielsen, and E. Marsh, "Speech in noisy environments (SPINE) adds new dimension to speech recognition R&D," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 212–216, 2002.

[59] R. Cutler, Y. Rui, A. Gupta, J. J. Cadiz, I. Tashev, L.-W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *Proceedings of the ACM International Conference on Multimedia*, pp. 503–512, 2002.

[60] P. Dai, U. Iurgel, and G. Rigoll, "A novel feature combination approach for spoken document classification with support vector machines," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR) Multimedia Information Retrieval Workshop*, 2003.

[61] F. M. G. de Jong, D. W. Oard, W. F. L. Heeren, and R. J. F. Ordelman, "Access to recorded interviews: A research agenda," *ACM Journal on Computing and Cultural Heritage*, vol. 1, no. 1, pp. 3:1–3:27, 2008.

[62] F. M. G. de Jong, R. J. F. Ordelman, and M. A. H. Huijbregts, "Automated speech and audio analysis for semantic access to multimedia," in *Semantic Multimedia*, vol. 4306 of *Lecture Notes in Computer Science,* Chapter 18, (Y. Avrithis, Y. Kompatsiaris, S. Staab, and N. O'Connor, eds.), pp. 226–240, Springer Berlin/Heidelberg: Berlin, Heidelberg, 2006.

[63] F. M. G. de Jong, T. Westerveld, and A. P. de Vries, "Multimedia search without visual analysis: The value of linguistic and contextual information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 365–371, 2007.

[64] S. Deerwester, "Improving information retrieval with latent semantic indexing," in *Proceedings of the 51st ASIS Annual Meeting*, vol. 25, (C. L. Borgman and E. Y. H. Pai, eds.), 1988.

[65] A. Désilets, B. de Bruijn, and J. Martin, "Extracting keyphrases from spoken audio documents," in *Information Retrieval Techniques for Speech Applications*, pp. 36–50, London, UK, Springer, 2002.

[66] G. Dias, E. Alves, and J. G. P. Lopes, "Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation," in *Proceedings of the National Conference on Artificial Intelligence — Volume 2*, pp. 1334–1339, 2007.

[67] R. M. W. Dixon, *The Rise and Fall of Languages.* Cambridge University Press, 1998.

[68] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/221–I/224, 1995.

[69] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[70] M. Fapšo, P. Smrž, P. Schwarz, I. Szöke, J. Schwarz, , M. Černocký, M. Karafiát, and L. Burget, "Information retrieval from spoken documents," in *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 410–416, 2006.

[71] M. Federico, "A system for the retrieval of Italian broadcast news," *Speech Communication*, vol. 32, no. 1–2, pp. 37–47, 2000.

[72] A. Ferrieux and S. Peillon, "Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, 1999.

[73] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover)," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–352, 1997.

[74] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, (R. Stiefelhagen, R. Bowers, and J. G. Fiscus, eds.), pp. 373–389, Berlin/Heidelberg: Springer-Verlag, 2008.

[75] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), Searching Spontaneous Conversational Speech Workshop*, pp. 45–50, Amsterdam, Netherlands, 2007.

[76] J. T. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[77] J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young, "Talker-independent keyword spotting for information retrieval," in *Proceedings of Eurospeech*, pp. 2145–2148, 1995.

[78] M. Fuller, M. Tsagkias, E. Newman, J. Besser, M. Larson, G. J. F. Jones, and M. de Rijke, "Using term clouds to represent segment-level semantic content of

podcasts," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), Searching Spontaneous Conversational Speech Workshop*, 2008.

[79] S. Furui and T. Kawahara, "Transcription and distillation of spontaneous speech," in *Springer Handbook of Speech Processing*, Chapter 32, (J. Benesty, M. M. Sondhi, and Y. A. Huang, eds.), pp. 627–652, Berlin/Heidelberg: Springer Berlin/Heidelberg, 2008.

[80] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[81] M. Gales and S. J. Young, *The Application of Hidden Markov Models in Speech Recognition*. now Publishers Inc., February 2008.

[82] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the RIAO Conference on Content-Based Multimedia Information Access*, (J.-J. Mariani and D. Harman, eds.), pp. 1–20, 2000.

[83] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, and V. M. Stanford, "Spoken document retrieval: 1998 evaluation and investigation of new metrics," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, pp. 1–7, 1999.

[84] J.-L. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," *Communications of the ACM*, vol. 13, no. 2, pp. 64–70, 2000.

[85] L. Gillick, J. Baker, J. Bridle, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, R. Roth, and F. Scattone, "Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech," in *Proceedings of the IEEE International Conference on Acoustics Speech, and Signal Processing*, pp. II/471–II474, 1993.

[86] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzila, "Recent progress in the MIT spoken lecture processing project," in *Proceedings of Interspeech*, pp. 2556–2556, 2007.

[87] U. Glavitsch and P. Schäuble, "A system for retrieving speech documents," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 168–176, 1992.

[88] U. Glavitsch, P. Schäuble, and M. Wechsler, "Metadata for integrating speech documents in a text retrieval system," *SIGMOD Record*, vol. 23, no. 4, pp. 57–63, December 1994.

[89] A. Goker, J. Davies, and M. Graham, *Information Retrieval: Searching in the 21st Century*. John Wiley & Sons, 2007.

[90] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, Inc., 1999.

[91] J. Goldman, S. Renals, S. G. Bird, F. M. G. de Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. W. Oard, C. Stewart, and R. Wright, "Accessing the spoken word," *International Journal on Digital Libraries*, vol. 5, no. 4, pp. 287–298, 2005.

[92] M. Goto and J. Ogata, "PodCastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions," in *Proceedings of Interspeech*, pp. 3073–3076, 2011.

[93] M. Goto, J. Ogata, and K. Eto, "PodCastle: A Web 2.0 approach to speech recognition research," in *Proceedings of Interspeech*, pp. 2397–2400, 2007.

[94] S. Gustman, D. Soergel, D. Oard, W. Byrne, M. Picheny, B. Ramabhadran, and D. Greenberg, "Supporting access to large digital oral history archives," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 18–27, 2002.

[95] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young, "Segment generation and clustering in the HTK Broadcast News Transcription System," in *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 133–137, 1998.

[96] T. Hain, P. C. Woodland, G. Evermann, M. J. F. Gales, X. Liu, G. L. Moore, D. Povey, and L. Wang, "Automatic transcription of conversational telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1173–1185, 2005.

[97] D. Hakkani-Tür, F. Bechet, G. Riccardi, and G. Tür, "Beyond ASR 1-best: Using word confusion networks in spoken language understanding," *Computer Speech and Language*, vol. 20, no. 4, pp. 495–514, 2006.

[98] D. Hakkani-Tür and G. Riccardi, "A general algorithm for word graph matrix decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/596–I/599, 2003.

[99] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[100] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 712–730, 2005.

[101] A. Haubold, "Selection and ranking of text from highly imperfect transcripts for retrieval of video content," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 791–792, 2007.

[102] A. G. Hauptmann, "Speech recognition in the Informedia Digital Video Library: Uses and limitations," in *Proceedings of the International Conference on Tools with Artificial Intelligence*, p. 288, 1995.

[103] A. G. Hauptmann and M. G. Christel, "Successful approaches in the TREC video retrieval evaluations," in *Proceedings of the Annual ACM International Conference on Multimedia*, pp. 668–675, 2004.

[104] A. G. Hauptmann and H. Wactlar, "Indexing and search of multimodal information," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/195–I/198, 1997.

[105] A. G. Hauptmann and M. J. Witbrock, "Informedia: News-on-demand multimedia information acquisition and retrieval," in *Intelligent Multimedia Information Retrieval*, (M. T. Maybury, ed.), pp. 215–239, The MIT Press, 1997.

[106] M. A. Hearst, "Multi-paragraph segmentation of expository text," in *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pp. 9–16, 1994.

[107] M. A. Hearst, *Search User Interfaces.* Cambridge University Press, 2009.

[108] W. F. L. Heeren and F. M. G. de Jong, "Disclosing spoken culture: User interfaces for access to spoken word archives," in *Proceedings of the British HCI Group Annual Conference on Human Computer Interaction*, pp. 23–32, 2008.

[109] W. F. L. Heeren, L. van der Werff, R. J. F. Ordelman, A. van Hessen, and F. M. G. de Jong, "Radio Oranje: Searching the Queen's speech(es)," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, p. 903, 2007.

[110] I. L. Hetherington and V. W. Zue, "New words: Implications for continuous speech recognition," in *Proceedings of Eurospeech*, pp. 2121–2124, 1993.

[111] D. Hiemstra, "Using language models for information retrieval," PhD thesis, University of Twente, 2001.

[112] J. Hirschberg and S. Whittaker, "Studying search and archiving in a real audio database," in *Working Notes of the AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, pp. 70–76, 1997.

[113] J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira, and A. Singhal, "Finding information in audio: A new paradigm for audio browsing/retrieval," in *Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio*, pp. 117–122, 1999.

[114] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. IV/73–IV/76, 2007.

[115] Y.-C. Hsieh, Y.-T. Huang, C.-C. Wang, and L.-S. Lee, "Improved spoken document retrieval with dynamic key term lexicon and Probabilistic Latent Semantic Analysis (PLSA)," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/961–I/964, 2006.

[116] P.-Y. Hsueh and J. D. Moore, "Automatic topic segmentation and labeling in multiparty dialogue," in *IEEE Spoken Language Technology Workshop*, pp. 98–101, 2006.

[117] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

[118] M. A. H. Huijbregts, D. A. Leeuwen, and F. M. G. Jong, "The majority wins: A method for combining speaker diarization systems," in *Proceedings of Interspeech*, pp. 924–927, 2009.

[119] A. Jaimes, H. Bourlard, S. Renals, and J. Carletta, "Recording, summarizing, and accessing meeting videos: An overview of the AMI project," in *Proceeings of the IEEE International Conference of Image Analysis and Processing Workshops*, pp. 59–64, 2007.

[120] D. A. James, "A system for unrestricted topic retrieval from radio news broadcasts," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/279–I/282, 1996.

[121] D. A. James, "The application of classical information retrieval techniques to spoken documents," PhD Thesis, University of Cambridge, June 1995.

[122] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent wordspotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/377–I/380, 1994.

[123] A. Janin, L. Gottlieb, and G. Friedland, "Joke-o-Mat HD: Browsing sitcoms with human derived transcripts," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1591–1594, 2010.

[124] F. Jelinek, *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press, 1998.

[125] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[126] R. Jin and A. G. Hauptmann, "Automatic title generation for spoken broadcast news," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 1–3, 2001.

[127] S. E. Johnson, P. Jourlin, K. S. Jones, and P. Woodland, "Spoken document retrieval for TREC-9 at Cambridge University," in *Proceedings of the Text REtrieval Conference*, (E. Voorhees and D. Harman, eds.), pp. 117–126, 2000.

[128] G. J. F. Jones, "Exploring the incorporation of acoustic information into term weights for spoken document retrieval," in *Proceedings of the BCS Information Retrieval Specialist Group Colloquium on Information Retrieval Research*, pp. 118–131, 2000.

[129] G. J. F. Jones and C. H. Chan, "Multimedia information extraction," *Chapter Affect-Based Indexing for Multimedia Data*. IEEE Computer Society Press, 2012.

[130] G. J. F. Jones and R. Edens, "Automated alignment and annotation of audio-visual presentations," in *Research and Advanced Technology for Digital Libraries*, vol. 2458 of *Lecture Notes in Computer Science*, (M. Agosti and C. Thanos, eds.), pp. 187–196, Springer Berlin/Heidelberg, 2002.

[131] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young, "Video mail retrieval: The effect of word spotting accuracy on precision," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/309–I/312, 1995.

[132] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 30–38, 1996.

[133] G. J. F. Jones and D. A. James, "A critical review of state-of-the-art technologies for cross-language speech retrieval," in *Cross-Language Text and Speech Retrieval Papers from the 1997 AAAI Spring Symposium, Technical Report SS-97-05*, Menlo Park, California, 1997.

[134] G. J. F. Jones and A. M. Lam-Adesina, "Exeter at CLEF 2003: Cross-language spoken document retrieval experiments," in *Comparative Evaluation of Multilingual Information Access Systems*, vol. 3237 of *Lecture Notes in Computer Science*, (C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, eds.), pp. 553–558, Springer Berlin/Heidelberg, 2004.

[135] G. J. F. Jones, K. Zhang, E. Newman, and A. M. Lam-Adesina, "Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR) Searching Spontaneous Conversational Speech Workshop*, 2007.

[136] P. Jourlin, S. E. Johnson, K. Spärck Jones, and P. C. Woodland, "Improving retrieval on imperfect speech transcriptions (poster abstract)," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 283–284, 1999.

[137] P. Jourlin, S. E. Johnson, K. Spärk Jones, and P. C. Woodland, "Spoken document representations for probabilistic retrieval," *Speech Communication*, vol. 32, pp. 21–36, 2000.

[138] B. H. Juang and L. R. Rabiner, "Automatic speech recognition — a brief history of the technology," in *Elsevier Encyclopedia of Language and Linguistics,* Second Edition, Elsevier, 2005.

[139] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 2008.

[140] V. Kalnikaité and S. Whittaker, "Social summarization: Does social feedback improve access to speech data?," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 9–12, 2008.

[141] S. Kazemian, F. Rudzicz, G. Penn, and C. Munteanu, "A critical assessment of spoken utterance retrieval through approximate lattice representations," in *Proceeding of the ACM International Conference on Multimedia Information Retrieval*, pp. 83–88, 2008.

[142] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proceedings of Eurospeech*, pp. 827–830, 1997.

[143] J. Kilgour, J. Carletta, and S. Renals, "The ambient spotlight: Queryless desktop search from meeting speech," in *Proceedings of the ACM Multimedia Searching Spontaneous Conversational Speech Workshop*, pp. 49–52, 2010.

[144] W. Kim and J. Hansen, *Speechfind: Advances in Rich Content Based Spoken Document Retrieval*. pp. 173–187. Information Science Reference, 2009.

[145] D. G. Kimber, L. D. Wilcox, F. R. Chen, and T. P. Moran, "Speaker segmentation for browsing recorded audio," in *Conference Companion on Human Factors in Computing Systems*, pp. 212–213, 1995.

[146] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[147] J. Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication*, vol. 35, no. 1–2, pp. 21–30, 2001.

[148] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 61–69, 2005.

[149] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, vol. 2, no. 1, pp. 1–24, 2005.

[150] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'Ready: A meeting recorder and browser," *ACM Computing Surveyes*, vol. 1, no. 2, 1999.

[151] J. Kupiec, D. Kimber, and V. Balasubramanian, "Speech-based retrieval using semantic co-occurrence filtering," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 350–354, 1994.

[152] M. Kurimo, "Thematic indexing of spoken documents by using self-organizing maps," *Speech Communication*, vol. 38, no. 1–2, pp. 29–45, 2002.

[153] M. Kurimo and V. Turunen, "An evaluation of a spoken document retrieval baseline system in finnish," in *Proceedings of Interspeech*, pp. 1585–1588, 2004.

[154] A. M. Lam-Adesina and G. J. F. Jones, "Using string comparison in context for improved relevance feedback in different text media," in *Proceedings of the String Processing on Information Retrieval Conference*, pp. 229–241, 2006.

[155] M. Larson and S. Eickeler, "Using syllable-based indexing features and language models to improve German spoken document retrieval," in *Proceedings of Interspeech*, pp. 1217–1220, 2003.

[156] M. Larson, M. Eskevich, R. J. F. Ordelman, C. Kofler, S. Schmiedeke, and G. J. F. Jones, "Overview of MediaEval 2011 rich speech retrieval task and genre tagging task," in *Working Notes Proceedings of the MediaEval Workshop*, CEUR-WS.org, 2011.

[157] M. Larson and J. Köhler, "Structured audio player: Supporting radio archive workflows with automatically generated structure metadata," in *Proceedings of the RIAO Conference on Large-scale Semantic Access to Content (Text, Image, Video and Sound)*, 2007.

[158] M. Larson, E. Newman, and G. J. F. Jones, "Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content," in *Proceedings of the Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, (C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, A. Peñas, G. J. F. Jones, M. Kurimo, T. Mandl, and V. Petras, eds.), pp. 906–917, Springer Berlin/Heidelberg, 2009.

[159] M. Larson, E. Newman, and G. J. F. Jones, "Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment," in *Multilingual Information Access Evaluation II. Multimedia Experiments*, vol. 6242 of *Lecture Notes in Computer Science*, (C. Peters, B. Caputo, J. Gonzalo, G. J. F. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsikrika, eds.), pp. 354–368, Springer Berlin/Heidelberg, 2010.

[160] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. J. F. Jones, "The community and the crowd: Developing large-scale data collections for multimedia benchmarking," *IEEE Multimedia*, IEEE Computer Society Digital Library. IEEE Computer Society, 15 May 2012.

[161] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. a. Murdock, G. Friedland, R. J. F. Ordelman, and G. J. F. Jones, "Automatic tagging

and geotagging in video collections and communities," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pp. 1–51, 2011.

[162] M. Larson, M. Tsagkias, J. He, and M. de Rijke, "Investigating the global semantic impact of speech recognition error on spoken content collections," in *Advances in Information Retrieval. Proceedings of the European Conference on IR Research*, vol. 5478 of *Lecture Notes in Computer Science*, (M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, eds.), pp. 755–760, Springer Berlin/Heidelberg, 2009.

[163] J. Laver, *Principles of Phonetics (Cambridge Textbooks in Linguistics)*. Cambridge University Press, 1994.

[164] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 120–127, 2001.

[165] D. Lee and G. G. Lee, "A Korean spoken document retrieval system for lecture search," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR) Searching Spontaneous Conversational Speech Workshop*, 2008.

[166] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42–60, 2005.

[167] S.-W. Lee, K. Tanaka, and Y. Itoh, "Combining multiple subword representations for open-vocabulary spoken document retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 505–508, 2005.

[168] B. Liu and D. W. Oard, "One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 673–674, 2006.

[169] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[170] W.-K. Lo, H. Meng, and P. C. Ching, "Cross-language spoken document retrieval using HMM-based retrieval model with multi-scale fusion," *ACM Transactions on Asian Language Information Processing*, vol. 2, no. 1, pp. 1–26, 2003.

[171] J. Löffler, K. Biatov, C. Eckes, and J. Köhler, "IFINDER: An MPEG-7-based retrieval system for distributed multimedia content," in *Proceedings of the ACM International Conference on Multimedia*, pp. 431–435, 2002.

[172] B. Logan, P. Moreno, and O. Deshmukh, "Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 31–35, 2002.

[173] B. Logan and J. M. V. Thong, "Confusion-based query expansion for OOV words in spoken document retrieval," in *Proceedings of Interspeech*, pp. 1997–2000, 2002.

[174] B. Logan, J. M. Van Thong, and P. J. Moreno, "Approaches to reduce the effects of OOV queries on indexed spoken audio," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 899–906, 2005.

[175] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pp. 25–32, 2006.

[176] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 51–58, 2006.

[177] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimisation," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[178] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[179] J. Mauclair, Y. Estève, S. Petitrenaud, and P. Deléglise, "Automatic detection of well recognized words in automatic speech transcription," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2006.

[180] M. T. Maybury, ed., *Intelligent Multimedia Information Retrieval*. The MIT Press, 1997.

[181] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek, "Approaches to topic identification on the switchboard corpus," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/385–I/388, 1994.

[182] C.-H. Meng, H.-Y. Lee, and L.-S. Lee, "Improved lattice-based spoken document retrieval by directly learning from the evaluation measures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4893–4896, 2009.

[183] H. Meng, B. Chen, S. Khudanpur, G. Levow, W. Lo, D. W. Oard, P. Schone, K. Tang, H. Wang, and J. Wang, "Mandarin-English Information (MEI): Investigating translingual speech retrieval," *Computer Speech and Language*, vol. 18, no. 2, pp. 163–179, 2004.

[184] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for German spoken term detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4885–4888, 2009.

[185] T. Mertens, D. Schneider, and J. Köhler, "Merging search spaces for spoken term detection," in *Proceedings of Interspeech*, pp. 2127–2130, 2009.

[186] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 472–479, 2005.

[187] G. Mishne and M. de Rijke, "Boosting web retrieval through query operations," in *Advances in Information Retrieval*, pp. 502–516, Springer, 2005.

[188] J. Mizuno, J. Ogata, and M. Goto, "A similar content retrieval method for podcast episodes," in *IEEE Spoken Language Technology Workshop*, pp. 297–300, 2009.

[189] L. L. Molgaard, K. W. Jorgensen, and L. K. Hansen, "Castsearch — context based spoken document retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. IV/93–IV/96, 2007.

[190] J. Morang, R. J. F. Ordelman, F. M. G. de Jong, and A. J. van Hessen, "Infolink: Analysis of dutch broadcast news and cross-media browsing," in *IEEE International Conference on Multimedia and Expo*, pp. 1582–1585, 2005.

[191] N. Moreau, S. Jin, and T. Sikora, "Comparison of different phone-based spoken document retrieval methods with text and spoken queries," in *Proceedings of Interspeech*, pp. 641–644, 2005.

[192] N. Moreau, H.-G. Kim, and T. Sikora, "Phonetic confusion based document expansion for spoken document retrieval," in *Proceedings of Interspeech*, pp. 1593–1596, 2004.

[193] P. J. Moreno, J. M. Van Thong, B. Logan, and G. J. F. Jones, "From multimedia retrieval to knowledge management," *Computer*, vol. 35, no. 4, pp. 58–66, 2002.

[194] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James, "The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives," in *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) Conference on Human Factors in Computing Systems*, pp. 493–502, 2006.

[195] C. Ng, R. Wilkinson, and J. Zobel, "Experiments in spoken document retrieval using phoneme n-grams," *Speech Communication*, vol. 32, no. 1–2, pp. 61–77, 2000.

[196] K. Ng and V. W. Zue, "Subword unit representations for spoken document retrieval," in *Proceedings of Eurospeech*, pp. 1607–1610, 1997.

[197] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[198] T. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Communication*, vol. 49, no. 6, pp. 453–463, 2007.

[199] NIST, *The Spoken Term Detection (STD) 2006 Evaluation Plan*, 2006.

[200] J. Nouza, J. Žďánský, P. Červa, and J. Kolorenč, "A system for information retrieval from large records of Czech spoken data," in *Text, Speech and Dialogue*, vol. 4188 of *Lecture Notes in Computer Science*, (P. Sojka, I. Kopeček, and K. Pala, eds.), pp. 485–492, Springer Berlin/Heidelberg, 2006.

[201] P. Nowell and R. K. Moore, "The application of dynamic programming techniques to non-word based topic spotting," in *Proceedings of Eurospeech*, pp. 1355–1358, 1995.

[202] D. W. Oard, "Speech-based information retrieval for digital libraries," Technical Report CS-TR-3778, University of Maryland, 1997.

[203] D. W. Oard, "User interface design for speech-based retrieval," *Bulletin of the American Society for Information Science and Technology*, vol. 26, no. 5, pp. 20–22, 2000.

[204] D. W. Oard, D. Soergel, D. Doermann, X. Huang, C. G. Murray, J. Wang, B. Ramabhadran, M. Franz, S. Gustman, J. Mayfield, L. Kharevych, and S. Strassel, "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 41–48, 2004.

[205] D. W. Oard, J. Wang, G. J. F. Jones, R. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran, "Overview of the CLEF-2006 cross-language speech retrieval track," in *Evaluation of Multilingual and Multi-modal Information Retrieval*, vol. 4730 of *Lecture Notes in Computer Science*, (C. Peters, P. Clough, F. Gey, J. Karlgren, B. Magnini, D. Oard, M. de Rijke, and M. Stempfhuber, eds.), pp. 744–758, Springer Berlin/Heidelberg, 2007.

[206] N. A. O'Connor, H. Lee, A. F. Smeaton, G. J. F. Jones, E. Cooke, H. Le Borgne, and C. Gurrin, "Fischlar-TRECVid-2004: Combined text- and image-based searching of video archives," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2006.

[207] J. Ogata, M. Goto, and K. Eto, "Automatic transcription for a Web 2.0 service to search podcasts," in *Proceedings of Interspeech*, pp. 2617–2620, 2007.

[208] J. S. Olsson, "Vocabulary independent discriminative term frequency estimation," in *Proceedings of Interspeech*, pp. 2187–2190, 2008.

[209] J. S. Olsson and D. W. Oard, "Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 91–98, 2009.

[210] J. S. Olsson and D. W. Oard, "Phrase-based query degradation modeling for vocabulary-independent ranked utterance retrieval," in *Proceedings of Human Language Technologies Conferemce of the North American Chapter of the Association for Computational Linguistics*, pp. 182–190, 2009.

[211] R. J. F. Ordelman, W. F. L. Heeren, M. A. H. Huijbregts, F. M. G. de Jong, and D. Hiemstra, "Towards affordable disclosure of spoken heritage archives," *Journal of Digital Information, Special Issue on Information Access to Cultural Heritage*, vol. 10, no. 6, 2009.

[212] R. J. F. Ordelman, A. J. van Hessen, and F. M. G. de Jong, "Speech recognition issues for Dutch spoken document retrieval," in *Proceedings of the International Conference on Text, Speech and Dialogue*, pp. 258–265, 2001.

[213] G. Paaß, E. Leopold, M. Larson, J. Kindermann, and S. Eickeler, "SVM classification using sequences of phonemes and syllables," in *Principles of Data Mining and Knowledge Discovery*, vol. 2431 of *Lecture Notes in Computer Science*, (T. Elomaa, H. Mannila, and H. Toivonen, eds.), pp. 373–384, Springer Berlin/Heidelberg, 2002.

[214] D. S. Pallett, J. S. Garofolo, and J. G. Fiscus, "Measurements in support of research accomplishments," *Communications of the ACM*, vol. 43, no. 2, pp. 75–79, 2000.

[215] Y.-C. Pan and L.-S. Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *IEEE Transactions on Speech and Audio Processing*, vol. 18, no. 6, pp. 1562–1574, 2010.

[216] S. Páraic, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 99–108, 1997.

[217] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard, "Overview of the CLEF 2007 cross-language speech retrieval track," in *Advances in Multilingual and Multimodal Information Retrieval*, vol. 5152 of *Lecture Notes in Computer Science*, (C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos, eds.), pp. 674–686, Springer Berlin/Heidelberg, 2008.

[218] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 275–281, 1998.

[219] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1973–1976, 2009.

[220] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[221] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[222] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[223] D. R. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[224] G. Rigoll, "The ALERT system: Advanced broadcast speech recognition technology for selective dissemination of multimedia Information," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 301–306, 2001.

[225] S. E. Robertson, "On term selection for query expansion," *Journal of Documentation*, vol. 46, no. 4, pp. 359–364, 1990.

[226] S. E. Robertson and K. Spärk Jones, "Relevance weighting of search terms," *Journal of the American Society of Information Science*, vol. 27, no. 3, pp. 129–146, 1976.

[227] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[228] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Text REtrieval Conference*, pp. 109–126, 1996.

[229] S. E. Robertson, H. Zaragoza, and M. J. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the International Conference on Information and Knowledge Management*, pp. 42–49, 2004.

[230] R. C. Rose, “Techniques for information retrieval from speech messages,” *Lincoln Laboratory Journal*, vol. 4, no. 1, pp. 45–60, 1991.

[231] R. C. Rose, E. I. Chang, and R. P. Lippmann, “Techniques for information retrieval from voice messages,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/317–I/320, 1991.

[232] R. C. Rose and D. B. Paul, “A hidden Markov model based keyword recognition system,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/129–I/132, 1990.

[233] S. Rosset, O. Galibert, G. Adda, and E. Bilinski, “The LIMSI QAst systems: Comparison between human and automatic rules generation for question-answering on speech transcriptions,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 647–652, 2007.

[234] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for TV baseball programs,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 105–115, 2000.

[235] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[236] M. Sanderson and F. Crestani, “Mixing and merging for spoken document retrieval,” in *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries*, pp. 397–407, 1998.

[237] M. Sanderson and X.-M. Shou, “Search of spoken documents retrieves well recognized transcripts,” in *Advances in Information Retrieval. Proceedings of the European Conference on IR Research*, (G. Amati, C. Carpineto, and G. Romano, eds.), pp. 505–516, Springer Berlin/Heidelberg, 2007.

[238] M. Saraclar and R. W. Sproat, “Lattice-based search for spoken utterance retrieval,” in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 129–136, 2004.

[239] T. Schaaf and T. Kemp, “Confidence measures for spontaneous speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II/875–II/878, 1997.

[240] P. Schäuble and U. Glavitsch, “Assessing the retrieval effectiveness of a speech retrieval system by simulating recognition errors,” in *Proceedings of the Workshop on Human Language Technology*, pp. 347–349, 1994.

[241] P. Schäuble and M. Wechsler, “First experiences with a system for content based retrieval of information from speech recordings,” in *Proceedings of the IJCAI Workshop on Intelligent Multimedia Information Retrieval*, pp. 59–69, 1995.

[242] C. Schmandt, “The intelligent ear: A graphical interface to digital audio,” in *Proceedings of the Internationl Conference on Cybernetics and Society*, pp. 393–397, 1981.

[243] D. Schneider, “Holistic vocabulary independent spoken term detection,” PhD thesis, University of Bonn, 2011.

[244] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector

machine-belief network architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/577–I/580, 2004.

[245] A. Siegler, M. A. amd Berger, M. Witbrock, and A. Hauptmann, "Experiments in spoken document retrieval at CMU," in *Proceedings of the Text Retrieval Conference*, pp. 319–326, 1998.

[246] M. Siegler and M. Witbrock, "Improving the suitability of imperfect transcriptions for information retrieval from spoken documents," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/505–I/508, 1999.

[247] M. A. Siegler, "Integration of continuous speech recognition and information retrieval for mutually optimal performance," PhD thesis, Carnegie Mellon University, 1999.

[248] J. Silva, C. Chelba, and A. Acero, "Integration of metadata in spoken document search using position specific posterior latices," in *Proceedings of the IEEE Spoken Language Technology Workshop*, pp. 46–49, 2006.

[249] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 21–29, 1996.

[250] A. Singhal, J. Choi, D. Hindle, D. D. Lewis, and F. Pereira, "AT&T at TREC-7," in *Proceedings of the Text REtrieval Conference*, pp. 239–252, 1999.

[251] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 34–41, 1999.

[252] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Proceedings of Interspeech*, pp. 53–56, 2005.

[253] L. Slaughter, D. W. Oard, V. L. Warnick, J. L. Harding, and G. J. Wilkerson, "A graphical interface for speech-based retrieval," in *Proceedings of the ACM Conference on Digital Libraries*, pp. 305–306, 1998.

[254] A. F. Smeaton, M. Morony, G. Quinn, and R. Scaife, "Taiscéalaí: Information retrieval from an archive of spoken radio news," in *Research and Advanced Technology for Digital Libraries*, vol. 1513 of *Lecture Notes in Computer Science*, (C. Nikolaou and C. Stephanidis, eds.), pp. 429–442, Springer Berlin/Heidelberg, 1998.

[255] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[256] K. Spärck Jones, G. J. F. Jones, J. T. Foote, and S. J. Young, "Experiments in spoken document retrieval," *Information Processing and Management*, vol. 32, no. 4, pp. 399–417, 1996.

[257] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 81–87, 2000.

[258] L. A. Stark, S. Whittaker, and J. Hirschberg, "ASR satisficing: The effects of ASR accuracy on speech retrieval," in *Proceedings of Interspeech*, pp. 1069–1072, 2000.

[259] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[260] A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, and K. Sönmez, "Combining words and speech prosody for automatic topic segmentation," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 61–64, 1999.

[261] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication*, vol. 29, no. 2–4, pp. 225–246, 1999.

[262] J. Tejedor, M. Fapso, I. Szoke, J. Cernocky, and F. Grezl, "Comparison of methods for language-dependent and language-independent Query-by-Example spoken term detection," *ACM Transactions on Information Systems*, vol. 30, no. 3, 2012.

[263] J. Tejedor, D. Wang, J. Frankel, S. King, and J. Colas, "A comparison of grapheme and phoneme-based units for Spanish spoken term detection," *Speech Communication*, vol. 50, no. 11–12, pp. 980–991, 2008.

[264] K. Thambiratnam and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 346–357, 2007.

[265] T. Tombros and F. Crestani, "A study of users' perception of relevance of spoken documents," Technical Report TR-99-013, International Computer Science Institute, 1999.

[266] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[267] B. T. Truong, S. Venkatesh, and C. Dorai, "Automatic genre identification for content-based video categorization," in *Proceedings of the International Conference on Pattern Recognition*, vol. 4, pp. 230–233, 2000.

[268] M. Tsagkias, M. Larson, and M. de Rijke, "Term clouds as surrogates for user generated speech," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 773–774, 2008.

[269] S. Tucker, N. Kyprianou, and S. Whittaker, "Time-compressing speech: ASR transcripts are an effective way to support gist extraction," in *Machine Learning for Multimodal Interaction*, vol. 5237 of *Lecture Notes in Computer Science* Chapter 21, (A. Popescu-Belis and R. Stiefelhagen, eds.), pp. 226–235, Springer Berlin/Heidelberg, 2008.

[270] S. Tucker and S. Whittaker, "Temporal compression of speech: An evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, 2008.

[271] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 631–638, 2007.

[272] A. van den Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 45–53, 1993.

[273] C. J. van Rijsbergen, *Information Retrieval.* Butterworths, 1979.

[274] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[275] M. Viswanathan, H. S. M. Beigi, S. Dharanipragada, and A. Tritschler, "Retrieval from spoken documents using content and speaker information," in *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 567–572, 1999.

[276] C. C. Vogt and G. W. Cottrell, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–173, 1999.

[277] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval.* Digital Libraries and Electronic Publishing, The MIT Press, 2005.

[278] H. D. Wactlar, A. G. Hauptmann, M. G. Christel, R. A. Houghton, and A. M. Olligschlaeger, "Complementary video and audio analysis for broadcast news archives," *Communications of the ACM*, vol. 43, no. 2, pp. 42–47, 2000.

[279] A. Waibel and K.-F. Lee, eds., *Readings in Speech Recognition.* Morgan Kaufmann, 1990.

[280] D. Wang, "Out-of-vocabulary spoken term detection," PhD thesis, University of Edinburgh, 2009.

[281] D. Wang, S. King, J. Frankel, R. Vipperla, N. Evans, and R. Troncy, "Direct posterior confidence estimation for out-of-vocabulary spoken term detection," *ACM Transactions on Information System*, vol. 30, no. 3, 2012.

[282] H.-M. Wang, "Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese," *Speech Communuication*, vol. 32, no. 1–2, pp. 49–60, 2000.

[283] H.-M. Wang, "Mandarin spoken document retrieval based on syllable lattice matching," *Pattern Recognition Letters*, vol. 21, no. 6–7, pp. 615–624, 2000.

[284] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 577–582, 2003.

[285] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 28–38, 2008.

[286] V. Warnke, S. Harbeck, E. Noth, and H. Niemann, "Topic spotting using subword units," in *9. Aachener Kolloqium "Signaltheorie" Bild- und Sprachsignale*, pp. 287–291, 1997.

[287] C. L. Wayne, "Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2000.

[288] M. Wechsler, E. Munteanu, and P. Schäuble, "New techniques for open-vocabulary spoken document retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 20–27, 1998.

[289] M. Wechsler, E. Munteanu, and P. Schäuble, "New approaches to spoken document retrieval," *Information Retrieval*, vol. 3, no. 3, pp. 173–188, 2000.

[290] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/297–I/300, 1995.

[291] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 16–19, 1996.

[292] P. Wellner, M. Flynn, and M. Guillemot, "Browsing recorded meetings with ferret," in *Machine Learning for Multimodal Interaction*, vol. 3361 of *Lecture Notes in Computer Science*, (S. Bengio and H. Bourlard, eds.), pp. 12–21, Springer Berlin/Heidelberg, 2005.

[293] P. Wellner, M. Flynn, A. Tucker, and A. Whittaker, "A meeting browser evaluation test," in *Computer-Human Interaction Extended Abstracts on Human Factors in Computing Systems*, 2005.

[294] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[295] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X. Huang, "Overview of the CLEF-2005 cross-language speech retrieval track," in *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*, (C. Peters, F. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, eds.), pp. 744–759, Springer Berlin/Heidelberg, 2006.

[296] E. W. D. Whittaker, J. M. Van Thong, and P. J. Moreno, "Vocabulary independent speech recognition using particles," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 315–318, 2001.

[297] S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, L. Isenhour, P. Stead, G. Zamchick, and A. Rosenberg, "Scanmail: A voicemail interface that makes speech browsable readable and searchable," in *Proceedings of the Special Interest Group on Computer-Human Interaction (SIGCHI) Conference on Human Factors in Computing Systems*, pp. 275–282, 2002.

[298] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "SCAN: Designing and evaluating user interfaces to support retrieval from speech archives," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 26–33, 1999.

[299] S. Whittaker, S. Tucker, K. Swampillai, and R. Laban, "Design and evaluation of systems to support interaction capture and retrieval," *Personal Ubiquitous Computing*, vol. 12, no. 3, pp. 197–221, 2008.

[300] L. Wilcox, F. Chen, and V. Balasubramanian, "Segmentation of speech using speaker identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I/161–I/164, 1994.

[301] L. D. Wilcox and M. A. Bush., "HMM-based wordspotting for voice editing and indexing," in *Proceedings of Eurospeech*, pp. 25–28, 1991.

[302] D. Willett, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence measures for HMM-based speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 3241–3244, 1998.

[303] M. J. Witbrock and A. G. Hauptmann, "Speech recognition and information retrieval: Experiments in retrieving spoken documents," in *Proceedings of the DARPA Speech Recognition Workshop*, 1997.

[304] M. J. Witbrock and A. G. Hauptmann, "Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents," in *Proceedings of the ACM International Conference on Digital Libraries*, pp. 30–35, 1997.

[305] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann, 1999.

[306] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. Spärck Jones, "Effects of out of vocabulary words in spoken document retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pp. 372–374, 2000.

[307] B. Wrede and E. Shriberg, "Spotting "Hot Spots" in meetings: Human judgments and prosodic cues," in *Proceeindgs of Eurospeech*, pp. 2805–2808, 2003.

[308] C.-H. Wu, C.-L. Huang, W.-C. Lee, and Y.-S. Lai, "Speech-annotated photo retrieval using syllable-transformed patterns," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 6–9, 2009.

[309] H. Yan, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust faster than real-time speaker diarization," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 693–698, 2007.

[310] H. Yang, L. Chaisorn, Y. Zhao, S. Y. Neo, and T. S. Chua, "VideoQA: question answering on news video," in *Proceedings of the ACM International Conference on Multimedia*, pp. 632–641, 2003.

[311] S. R. Young, "Detecting misrecognitions and out-of-vocabulary words," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II/21–II/24, 1994.

[312] P. Yu, K. Chen, C. Ma, and F. Seide, "Vocabulary-independent indexing of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 635–643, 2005.

[313] T. Zhang and C. C. Kuo, *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing.* Kluwer Academic Publishers, 2001.

[314] T. Zhang and C.-C. J. Kuo, “Heuristic approach for generic audio data segmentation and annotation,” in *Proceedings of the ACM International Conference on Multimedia (Part 1)*, pp. 67–76, 1999.

[315] Z.-Y. Zhou, P. Yu, C. Chelba, and F. Seide, “Towards spoken-document retrieval for the internet: Lattice indexing for large-scale web-search architectures,” in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 415–422, 2006.

[316] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, “Speaker diarization: From broadcast news to lectures,” in *Machine Learning for Multimodal Interaction*, vol. 4299 of *Lecture Notes in Computer Science,* Chapter 35, (S. Renals, S. Bengio, and J. G. Fiscus, eds.), pp. 396–406, Springer Berlin/Heidelberg, 2006.

[317] G. Zweig, J. Makhoul, and A. Stolke, “Introduction to the special section on Rich Transcription,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1490–1491, 2006.