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

now
the essence of knowledge

Spoken Content Retrieval: A Survey of Techniques and Technologies

By Martha Larson and Gareth J. F. Jones

Contents

Foundations and Trends® in
Information Retrieval
Vol. 5, Nos. 4–5 (2011) 235–422

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.

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 time-point (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.

2

Overview of Spoken Content Indexing and Retrieval

In this module, we offer a high-level perspective on SCR that sets the stage for discussion of specific component technologies and issues in the material that follows. In subsection 2.1, we present an overview of the general architecture for a basic SCR system and briefly describe its various components. The overview provides a context for our presentation of background information on Information Retrieval (IR) techniques in subsection 2.2. Then, subsection 2.3 covers evaluation of IR systems. Finally, subsection 2.4 describes in detail those aspects of SCR that make it different from text-based IR, identifying a list of key challenges that are particular to SCR and are faced when designing and implementing an SCR system.

2.1 General Architecture of an SCR System

Although SCR systems are implemented differently depending on the deployment domain and the use scenario, the underlying architecture consists of a set of conventional components that remain more or less stable. This architecture is presented schematically in Figure 2.1, in

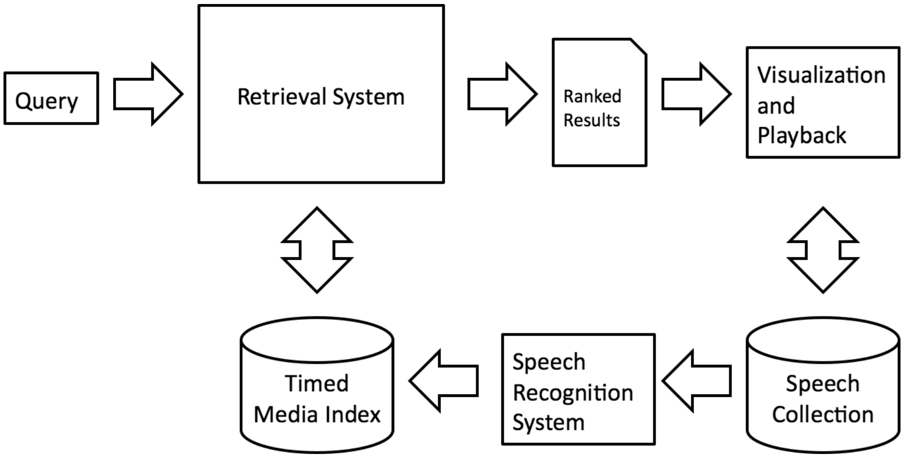


Fig. 2.1 Block diagram depicting an abstraction of a typical spoken content retrieval system.

order to present an initial impression of the technologies "under the hood" of a typical SCR system.

The *Query* depicted on the left represents the user input to the system. We emphasize that the query is not the actual information need of the user, but rather an attempt of the user to express this information need. Often the query is a highly under-specified representation of the information need, and part of the goal of the system will be to automatically enhance this specification in order to return useful results to the user.

The *Retrieval System* has the function of matching the query with the items in the collection. This matching takes place using one of the many IR frameworks that have been developed for text-based applications. These will be discussed further in subsection 2.2.

The retrieval system consults the index, here labeled *Timed Media Index*, which contains features that represent the items in the collections and, in general, also time-codes indicating the time points within each item associated with occurrences of these features. An index is a representation of a collection as indexing features, together with information that associates those features with particular items.

The process of indexing involves generation of an index, and can be defined as follows:

> **Spoken Content Indexing** is the task of generating representations of spoken content for use in a retrieval system. These representations include indexing features (i.e., terms consisting of words and phrases derived from the spoken content and also terms describing the spoken content, such as speaker identities), weights for the indexing terms and also time codes indicating the time points within the spoken content associated with the indexing terms.

The exact form of the index depends on the domain and the application. For example, for some applications, speech media is pre-segmented into documents and only entire items are returned to the user in the results list. In this case, the time code information returned by the speech recognizer may be discarded, and the index may contain only the information regarding which indexing term is associated with which document. In other cases, the system might return a time point within an item as a result. In this case, time code information must be retained in the index. The structure of the index is optimized for efficient calculation of matches between the query and the speech media item. Since indexing techniques are shared by SCR and IR, we will not treat them here, but rather refer the reader to the IR literature, for example, [57, 178, 305].

Indexing features are generated by the *Speech Recognition System* that processes the material in the *Speech Collection* at indexing time. The major source of features in the index is ASR, that is, the process that transcribes the spoken word to text. However, other sources of information, such as metadata and rich transcripts that include labels indicating who spoke when, are also important. The parts of this survey that are relevant to the processes of spoken content indexing are *Automatic Speech Recognition*, *Exploiting Automatic Speech Recognition Output*, and *Spoken Content Retrieval beyond ASR Transcripts*, which covers ways in which speech recognizer output can be exploited for indexing.

As the output of the SCR system, the user receives from the system a list of *Ranked Results*, a set of results ordered in terms of their likelihood of potential relevance to the query. A result can take the form of a spoken content item, a segment of the speech stream whose scope is dynamically customized to the query (sometimes referred to as a “relevance interval”) or a time point at which the user should start viewing/listening to the content (a so-called “listen-in point” or “jump-in point”). The choice of the form of results depends on the domain and on the use scenario. In any case, it is important that the results list contain proper surrogates, representations of each result that allow the user to make an initial judgment of whether or not the result is a good match to the information need without having to initiate playback.

Finally, the SCR system must offer the user a means of *Visualization and Playback* of the individual results. Result visualization in a playback interface is necessary for the same reason as surrogates are important for the results list: users must be able to judge the relevance of speech media results without listening to or viewing long swaths of audio or video content, a very time-consuming process. Time can also be saved by providing the user with an intelligent multimedia player, which makes it possible to jump directly to certain points within a speech media result, for example the beginning of a particular speaker turn. Issues concerning the user interface will be discussed in greater detail in *Accessing Information in Spoken Content*.

2.2 Information Retrieval for SCR

The *Retrieval System* module depicted in Figure 2.1 is most often taken to be a standard IR system, implementing one of the standard IR frameworks. Each of these frameworks matches items to queries, and outputs a ranked list of results. The difference between the frameworks lies in their choice of *ranking function*, which computes a *ranking score* (RS) for each of the result items. The RS is sometimes called the *Retrieval Status Value* (RSV). The RS is generated on the basis of indexing features that have been extracted from items. The standard IR frameworks all apply an independence assumption, also known as the *Bag of Words* approach, which means that the system disregards information about order or co-occurrence of features in a particular item and treats them as occurring independently of each other.

This survey does not review IR issues in depth. Instead, we offer an overview of existing IR models and refer the reader to the IR literature for full details of their theory, implementation, and evaluation [14, 57, 89, 178, 305]. The exposition is ordered in terms of the historical timeline along which the IR models were developed, which also reflects their relative complexity.

Boolean search. In the Boolean search framework, an item is returned as potentially relevant if its contents match a possibly very complex query constructed using Boolean operators. Since Boolean search does not generate a relevance score for each item, the set of returned results is not ordered by potential relevance. The searcher must either make use of other information, for example, date of creation, or abbreviated representation of their content (i.e., surrogates) to decide which items to inspect further.

A basic form of Boolean search query is a simple AND construction requiring all features in the query to present in the item in order for it to be retrieved. Taking a simple AND query, the Boolean search framework basically addresses a “finding mentions” task, looking for the presence of certain features in the items to be retrieved.

Vector Space Model. The Vector Space Model (VSM) makes use of a vector representation containing indexing features, $\vec{V}(i)$. One such representation is made for each item in the collection of available items. The elements of the vector contain the weights of the individual indexing features (i.e., terms). Each weight represents the importance of that feature for the item.

The VSM is based on the assumption that the closeness of vectors within the vector space reflects the semantic similarity of the items that they represent. Within the VSM, the ranking score of an item, with respect to the query, is calculated as the similarity between a vector representing the item and a vector representing the query. Query-document similarity in the VSM is calculated using the dot product between the vectors. This similarity can be used either without length normalization, as in Equation 2.1, or with the normalization, as in Equation 2.2.

$$\mathrm{RS}_{\mathrm{VSM}_{\mathrm{dot}}}(Q,I) = \mathrm{sim}_{\mathrm{dot}}(q,i) = \vec{V}(q) \cdot \vec{V}(i) \tag{2.1}$$

$$\mathrm{RS}_{\mathrm{VSM}_{\mathrm{cos}}}(Q,I) = \mathrm{sim}_{\mathrm{cos}}(q,i) = \frac{\vec{V}(q) \cdot \vec{V}(i)}{|\vec{V}(q)||\vec{V}(i)|} \tag{2.2}$$

There are several alternative schemes that can be used for calculating the weights of each term. Conventionally, these schemes make use of statistics calculated on the basis of term occurrences in the collection. The underlying principle that should be followed when choosing a weighting scheme is quite simple: the weighted terms for a certain item should be *representative* (i.e., capture the content of that item) and *discriminative* (i.e., capture properties of that item that make it different from other items).

A popular weighting scheme, referred to as *tf-idf* uses *tf*, term frequency (the number of occurrences of a given feature in the item), as the representative component and *idf*, the inverse document frequency (the inverse of the number of documents in the collection containing the given term), as the discriminative component. A detailed explanation of the VSM and presentation of alternative tf and *idf* component functions is given in [235]; this account is extended to incorporate a more effective method of query and document length normalization than in the standard cosine function in [249].

The VSM was the first widely used ranked IR framework, and has been used in a number of SCR studies. SCR work that makes use of the VSM for retrieval includes [196, 197, 241, 288, 304].

Probabilistic Retrieval. Probabilistic Retrieval (PR) is based on the Probability Ranking Principle, which states that the most effective retrieval system given the available data is the system that ranks items according to the probability of their relevance to the user's information need. In order to calculate the probability of an item being relevant to the information need, a particular model must be adopted. The common approach is to treat relevance as binary (i.e., an item is either relevant or it is not) and to apply Bayes' decision rule. Under this approach, an item is considered relevant if the probability that it belongs to the relevant class of items is larger than the probability that it belongs to the non-relevant class of items.

Analysis from this starting point, under the assumption that terms occur independently of each other, leads to the Binary Independence Model (BIM) [273]. Operationalizing this model for a particular query further leads to the Robertson/Spärck Jones relevance weight rw(t) of each term t [226] calculated as follows:

$$\mathrm{rw}(t) = \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(R - r_t + 0.5)(n_t - r_t + 0.5)}, \tag{2.3}$$

where R is the number of relevant documents for this query, r_t is the number of relevant documents containing term t, N is the total number of documents in the collection, and n is the total number of documents containing t. The addition of 0.5 facilitates estimation using small amounts of data — specifically, by avoiding the assumption that if a term has not appeared in any relevant documents so far, it will never appear in any.

In the absence of relevance information, rw(i) reduces to the *collection frequency weight* (cfw(t)),

$$\mathrm{cfw}(i) = \log \frac{N - n_t}{n_t}.$$

If, as is generally the case, $N \gg n_t$, this equation is a very close approximation to the standard *inverse document frequency* weight.

$$\mathrm{idf}(t) = \log \frac{N}{n_t}$$

Query-document similarity under the BIM is calculated as

$$\mathrm{RS}_{\mathrm{BIM}}(Q, I) = \sum_{t \in \mathrm{query}} \mathrm{cfw}(t). \tag{2.4}$$

This formula makes clear the relationship between the ranking score as calculated by the BIM for PR and the ranking score as calculated by the VSM in Equation 2.1. When *idf* term weights are used, the ranking score produced by a VSM making use of dot-product similarity is expressed as:

$$\mathrm{RS}_{\mathrm{VSM}} = \sum_{t \in \mathrm{query}} \log \frac{N}{n_t}. \tag{2.5}$$

Both models use a sum over individual term contributions because they both impose the assumption that terms are distributed independently of each other in items. Further, both models capture the way in which term occurrence serves to discriminate between documents.

The similarity between Equations 2.4 and 2.5 makes evident both the advantage and disadvantage of the BIM. The advantage is that the probabilistic framework achieves the same basic model as the VSM, known to work well in practice, but uses more principled means based on probability theory and a notion of relevance. The disadvantage is that the BIM captures no information about term frequencies. In other words, Equation 2.4 contains no term that is analogous to *tf* in the *tf*–*idf* weighting scheme for the VSM, which captures how representative a given term is for a given item. The lack of information about term frequencies compromises the performance of the BIM and led to the development of the extension of the BIM referred to as the Okapi or BM25 model [227, 228]. This extension introduces a sensitivity to term frequency and document length into the original BIM and has been demonstrated to yield good performance for many IR tasks including SCR [47, 137].

Language Modeling framework. The most recently introduced of the major IR frameworks is the Language Modeling (LM) framework [111, 218]. This framework takes a very different approach to the generation of a ranked list of potentially relevant items in response to a query. The model essentially estimates the probability of a language model associated with each item generating the query, or, alternatively, the query generating the item. This value is returned by the system as the ranking score. Conventionally, this problem is expressed using Bayes' Theorem and then simplified, as shown in Equation 2.6.

$$\mathrm{RS}_{\mathrm{LM}} = P(i|q) = \frac{P(q|i)P(i)}{P(q)} \propto P(q|i) \tag{2.6}$$

The simplification involves imposing the assumption that the prior probability of all items (i.e., the probability of items before the query is known) is the same, and noting that the prior probability of the query is the same for all documents so does not impact on the ranking of documents with respect to the ranking score. The model in Equation 2.6

is referred to as the query likelihood model since the language model used here is one that returns the likelihood of a query given a particular item. This likelihood is calculated via a language model of the document, as shown in Equation 2.7.

$$\mathrm{RS}_{\mathrm{LM}} = \sum_{w \in q} \log P(w|i) \tag{2.7}$$

Smoothing is used to prevent Equation 2.7 from returning a zero value if a query word is not contained in the document. A method often applied is to use a linear interpolation language model, as discussed by [178], that uses a linear combination between a multinomial distribution estimated for the item $P(w|c)$ and a multinomial distribution estimated over the background collection $P(w|c)$, as shown in Equation 2.8.

$$\mathrm{RS}_{\mathrm{LM}} = \sum_{w \in q} \lambda \log P(w|i) + 1 - \lambda \log P(w|c) \tag{2.8}$$

Estimation of the weighting parameter λ is important and is often carried out using a development data set containing queries and reference documents that have been hand labeled as relevant. As with the VSM and BIM discussed previously, this model makes use of the independence assumption. Since terms are assumed to have independent distributions, the contributions can be combined in a straightforward manner. In Equation 2.7, the contributions from terms are represented in the log domain to prevent underflow and they are combined using a simple sum. SCR work that has made use of the language modeling framework includes [47].

Relevance feedback and pseudo-relevance feedback. Relevance feedback in IR is a technique that aims to enrich the user's original query with further information concerning the user's underlying need. If the user identifies relevant items for the system, for example, by making a selection of items from the initial results list returned by the system, information from these items can be used as feedback to expand the query and revise parameters of the retrieval system. User-based relevance feedback can be approximated by making the assumption that the top ranked items in the initial results list are relevant and using these items as the feedback set. In this case, the technique is referred to as "pseudo-relevance feedback" (PRF) or "blind relevance feedback."

For text retrieval, the motivation to perform PRF is to create a query that gives a better specification of the user's information need with the aim of improving retrieval effectiveness. In the SCR setting, PRF has the potential to serve two purposes: expansion of the query to better describe the underlying information need, as in the case of text retrieval, but also to assist in addressing a particular problem of SCR. As mentioned in the *Introduction*, SCR effectiveness is impacted by the presence of errors in the ASR transcripts. In the SCR setting, PRF can expand the query to include terms that are well recognized by the ASR system, which can help to address retrieval problems arising when query terms are either missing from the vocabulary of the ASR system or are not well recognized by the ASR system.

Relevance feedback is implemented in different ways into the VSM, PR and LM IR frameworks. We give a brief description of each method here, and, again, make reference to IR texts for more detail.

Within the VSM framework, the general strategy is to adjust the vector representing the query away from non-relevant items within the vector space and towards relevant items. To this end, the original query vector is expanded using Rocchio's algorithm, which we give here in the formulation used by [178].

$$\vec{q_m} = \alpha\vec{q_0} + \beta\frac{1}{|D_r|}\sum_{\vec{d_j}\in D_r}\vec{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d_j}\in D_{nr}}\vec{d_j} \tag{2.9}$$

The new query, $\vec{q_m}$, consists of the original query, $\vec{q_0}$, weighted by a factor of α and modified with a positive contribution from the set of relevant documents, D_r, and a negative contribution from the set of non-relevant documents, D_{nr}. The contribution of each document set, weighted with a factor (β and γ), consists of the sum over the contributions of the individual vectors, $\vec{d_j}$, representing the documents in that set.

Within the PR framework, the approach is to seek to select expansion terms which when added to the query have the greatest utility in improving the rank of relevant documents in the ranked list. In order to do this, all terms appearing in relevant or assumed relevant documents are ranked using an offer weight ow(t) (sometimes referred to as the Robertson selection value rsv(t)). The derivation of ow(t) is described in [225], and is calculated as follows:

$$\mathrm{ow}(t) = r_t \times \mathrm{rw}(t), \tag{2.10}$$

where r_t and rw(t) have the same definitions used in the calculation of the PR term weights. Note that ow(t) represents a trade-off between the specificity of terms via rw(t) and the presence of a term in known relevant documents via r_t. Thus, terms achieving high ow(t) values are those with strong specificity, which occur in multiple relevant documents. The number of terms that should be added to the query must be determined empirically for a specific task.

Within the LM framework, PRF is implemented using a query model called a *relevance model*, which is a language model estimated to provide an enriched representation of the information need behind the query. The general strategy is to make use of information about which items are relevant in order to estimate a highly accurate relevance model that serves in place of the original query. Here, we present the formulation given by [57]. First, documents are ranked using the query likelihood score in Equation 2.7 and a set C of top-ranked documents is selected to form the relevance set, the basis for the relevance models. Next, for every word occurring in the relevance set, a relevance model probability $P(w|R)$ is calculated. Finally, new ranking scores are calculated for the documents using the negative cross-entropy of the relevance model and the original query likelihood model.

$$\mathrm{RS}_{\mathrm{RM}} = \sum_{w} P(w|R) \log P(w|D) \tag{2.11}$$

For details of the estimation of $P(w|R)$, we refer to [57, 164]. Relevance feedback has been shown to be an effective technique for improving SCR in many studies for a wide range of tasks. Further details on the use of relevance feedback in SCR are given in subsection 5.2.4.

Representing term dependence. The models that we have discussed thusfar have made use of the “Bag of Words” approach, that is, the assumption that terms within items are distributed independently of each other. The use of this assumption is widespread because it forms the basis for simple, yet effective systems. The utility of such systems is perhaps surprising due to the fact that the independence assumption represents a gross over-simplification of the reality of speech and language use. There are two factors that constrain the co-occurrence of words in human language: first, the requirements of syntactic structure, and second, the association between particular words and the concepts that humans speak about. The second factor is particularly important for IR, which regards words as the basic units that bear meaning within language. Certain meanings are expressed by using words in combination, and are absent if the same words are used individually.

The classic example is a phrasal noun such as "ground truth." When used together, the words provide strong evidence that a document is about a topic related to experimental evaluation. The evidence is stronger than twice the evidence contribution that would be made by an occurrence of "ground" without "truth" or an occurrence of "truth" without "ground."

A second example illustrates a more challenging case. Words whose co-occurrence is associated with specific topics are not necessarily adjacent within documents, rather they can be separated by an arbitrary number of intervening words. For example, the words "polish," "nails" and "filing" taken together constitute strong evidence that a particular document treats the topic of manicures. The combined evidence is clearly stronger that the individual contributions. A document containing the word "polish" could just as well concern gemstones, one containing "nails" could just as well be about carpentry and one containing the word "filing" could just as well be related to office work.

IR systems handle term dependence with various approaches. A straightforward approach to integrating proximity information is to use *phrase terms* (contiguous words) and *proximity terms* (words separated by a constrained number of intervening terms) alongside of conventional word-level terms [178]. Refer to [187] for details, in particular how weights are estimated for such terms for use in the VSM. A similar approach is the *dependence model*, which makes use of proximity terms within a framework that combines language modeling and an inference network [57, 186]. SCR research that has taken phrasal nouns and term proximity into account includes [38, 137].

Integrating additional information. Commercial search engines incorporate a wide range of features into the computation that generates their search results [57]. The features reach beyond the content of the results themselves. Perhaps the best known examples are techniques that exploit the hyperlinked structure of the collection, such as PageRank [24].

Research work on text-based IR provides us with various methods of integrating additional information into the retrieval framework. The simplest method is data fusion of retrieval results via Linear Combination of Scores (LCS) of retrieved item lists. This method can be used when more than one IR system or independent content index is available, each of which ranks items with respect to a different representation, for example, ASR transcript and metadata.

$$\mathrm{RS}_{\mathrm{LCS}} = \sum_{s \in \mathrm{systems}} w_s \mathrm{RS}_s(Q,I) \tag{2.12}$$

For a given document-query pair, the retrieval scores of each system $\mathrm{RS}_s(Q,I)$ are weighted with an empirically set weighting factor w_s and combined with a simple sum. LCS generally benefits retrieval effectiveness when the separate lists are of similar quality, but bring additional information to the retrieval ranking process [19, 276]. An early example of combining retrieval lists from multiple indexing sources is described in [132].

Another case arises when one of the information sources is basically query independent, such as, recency of the creation of the document. This feature can be integrated into the language modeling framework. In this case, the simplification in Equation 2.6 is made retaining the prior probability, i.e.,

$$\mathrm{RS}_{\mathrm{LM}} = P(i|q) = \frac{P(q|i)P(i)}{P(q)} \propto P(q|i)P(i). \tag{2.13}$$

The additional source of information can then be appropriately scaled and integrated into the retrieval score calculation as $P(i)$.

Other possibilities for integrating additional information into a retrieval system involve applying discriminative learning techniques to the ranking problem ("learning-to-rank"), applying reranking techniques, exploiting graph-based integration and making use of Bayesian Networks. Here, our intent is to point out the wide range of available approaches. We refer the reader to the IR literature for further information and recommend [57, 178] as starting points from which to explore the related literature.

Pre-processing. Pre-processing refers to the steps that are taken to modify the original text in order to make it more suited for indexing. Pre-processing for SCR follows similar methods to that for regular text retrieval, as explored in detail in most IR textbooks. For completeness, we briefly review common pre-precessing methods here, highlighting significant relevant points for SCR.

All pre-processing methods involve some degree of *normalization*, mapping lexical items (i.e., words) that are equivalent onto character strings that are completely identical. For example, words that are capitalized because they occur at the beginning of a sentence are often mapped to their lower case forms, so that both versions correspond to only a single term in the index. Some systems discard case altogether, which is an easy approach, but may eliminate information about valuable distinctions from the system, for example, between common and proper nouns. In some cases, documents will use different character encodings so it is important to make sure that the encoding is standardized and characters are not dropped or rendered unreadable in the process. Normalization can also involve standardizing spellings (e.g., U.K. vs. U.S. spelling standard for English) and dealing with spelling errors. Some systems map acronyms onto their full forms, (e.g., "F.C.C." is mapped to "Federal Communications Commission"). Normalization needs to be considered carefully for SCR. The output of ASR systems often does not provide casing as used in standard written texts, and spoken expressions are in some cases not of the same form as their written versions. The combination of ASR transcripts with metadata sources would require that their potentially rather different linguistic forms are normalized for consistency. However, it is clear that this mapping must be done correctly in order to avoid introducing errors into the text. For this reason, acronyms can also be treated elsewhere in an IR system on par with synonyms (e.g., “car” and “automobile”), for example during a query expansion step.

Often, an IR system applies some form of conflation of different forms of semantically related words. For English, these usually take the form of *stemming* [178]. As outlined in the *Introduction*, other languages often require different language specific processing, for example noun decompounding in German. Stemming is the process of mapping words onto their word stems or base forms. For example “tests,” “test,” “testing” and “tested” would all be mapped to the base form “test.” Stemming is useful because it collapses words that are semantically related, but slightly different in form, onto a single category. It should however, be applied with caution. For example, it is reported that users typically mean quite different things when entering the queries “apple” and “apples” into a search engine, the former referring to the company and the latter to the fruit.

In addition to stemming, *stopping* or *stopword removal* can also be applied. Stopwords are generally function words, such as the English words “the,” “a,” “her,” “that,” “for.” These words are removed since they have a syntactic function, but do not directly express meaning in the way that other words (known as *content words*) do. Their removal can produce significant computational savings in an IR system. Although the practice is common for IR systems, like stemming, it must be applied with discretion. Stopwords can contain important clues to the style or genre of the document (e.g., a personal testimony can be expected to have a high incidence of the terms “we” and “I”).

For some applications, these factors will contribute to relevance. For SCR, it is not always clear if applying stemming and stopping will enhance system effectiveness. Stemming could potentially have either a positive or negative effect. If the word in the original audio has the same base form as a word that the ASR system mis-recognizes in its place, then stemming can be beneficial. For example, if the original audio contained the sequence “test in” and the ASR system mis-recognized this sequence as “testing,” then stemming the word “testing” to its base form “test” would help to ameliorate the impact of the error. However, if the word in the original audio does not have the same base form as the mis-recognized word that replaces it, stemming could potentially conflate mis-recognized word forms, and possibly exacerbate the effect of word errors. For example, if "test in" in the original audio is recognized as "nesting," then stemming the word "nesting" to its base form "nest" could increase the impact of the word error on the SCR system. In some cases, stopword removal may be harmful. Stopwords in the ASR transcripts might not reflect actual spoken stopwords, but rather may be substituted for the actual spoken content words, which are either poorly articulated or outside the vocabulary of the ASR system. In this case, stopword patterns may offer a useful clue to the original spoken content and discarding them may remove potentially useful information from the SCR system.

Finally, we would like to mention the importance of *tokenization*, the process of mapping the characters occurring in the document to the words that will be used for indexing. Tokenization presents well understood challenges for written text content. While it is straightforward for a language like English, which uses white space to separate words, it is more challenging for a language like Chinese where text segmentation methods must be applied to identify word units. Related segmentation issues are raised in free compounding languages such as German, which, as mentioned previously, must often be split to identify their constituent words for indexing. ASR transcripts can present different tokenization challenges. For example, the recognizer can recognize the word "notebook" as the words "note" and "book." Thus, it can actually be useful to consider the mapping of some elements of ASR transcripts to compounds. If such confusions are frequent enough, as is likely the case with ASR systems that make use of vocabularies including sub words, recombining words can introduce a further source of error. In such cases, subwords may be used directly, as further discussed in *Exploiting Automatic Speech Recognition Output*.

2.3 Evaluation

The success of an IR system lies in its ability to satisfy user information needs. Since it is not always practical to carry out user studies to evaluate systems, evaluation is often accomplished using a system-oriented approach. Under such an approach, systems are evaluated by using a list of queries for which the relevant documents have been annotated by human assessors (called "the ground truth") and the system output is compared to this list. The accessor's annotation process serves to generate a representation of user perceptions of whether an item satisfies a particular information need. An advantage of the systems-oriented approach is that it creates highly controlled evaluation conditions that make it possible to compare performance across algorithms, systems and research sites.

In the field of IR, a variety of metrics are used to evaluate retrieval performance. The most familiar are precision and recall. Precision is defined as the proportion of retrieved documents that are relevant to the query. Recall is defined as the proportion of relevant documents that are retrieved. These measures are often reported for results lists of lengths $N = 5, 10, 20$, in which case they are labeled $P@N$ and $R@N$. Often, the F-measure, the harmonic mean of precision and recall, is also reported.

In order to express the quality of the overall results list (i.e., not only the top-n), Mean Average Precision (MAP) is commonly used. MAP is literally the *average* Average Precision (AP) — where AP is averaged across all queries in the query set. AP is calculated by moving down the results list for a given query and calculating precision at each rank N, where a document has been correctly retrieved. AP is then calculated by averaging $P@N$ over all N at which it has been calculated.

Another method of presenting a complete picture is to produce a precision-recall graph, which plots precision against recall for every point in the retrieved results list. Related is the ROC curve, which plots the precision with respect to the positive class against the precision with respect to the negative class.

An important dimension to IR evaluation is statistical significance, which measures the reliability in the comparison of experimental results. This topic is very important in the reporting of IR experimental results. An introduction to significance testing in IR is contained in [32].

Note that these measures are binary relevance judgments. In other words, a document is either judged relevant or not relevant. We refer the reader to the IR literature (e.g., [57, 178]) for more information on evaluation, and for an explanation of Normalized Discounted

Cumulative Gain (NDCG), which is used when relevance judgments are non-binary, that is, they have been made at a number of levels.

In the case of SCR, an evaluation metric adopted directly from IR is not always suitable. In particular, in cases in which spoken content is unstructured, there is no fundamental notion of a document over which precision or recall could be evaluated. An SCR system that returns "listen-in points" rather than items can be evaluated using the Generalized Average Precision (GAP) introduced by [168]. GAP is an extension of average precision that incorporates weights based on the distance between the "listen-in points" returned by the SCR system and the reference start parts (i.e., the manually labeled ground truth) [205].

2.4 Challenges for SCR

In the *Introduction*, we presented an initial overview of the basic differences between SCR and IR. Now that we have covered the basic IR frameworks, we shall revisit the issues in more detail. In this subsection, we discuss a list of areas in which SCR presents particular challenges, above and beyond the challenges of text IR. In order to design and develop an SCR system, it is important to go beyond a naïve combination of ASR and IR and to address these areas explicitly.

The challenge of handling uncertainty. The ASR systems that produce speech transcripts are also capable of generating some form of *confidence score*, values that reflect the level at which the recognizer is certain that the recognized word is indeed the word that was spoken in the speech signal. Conventionally, a confidence score reflects an acoustic match between the signal and the transcribed word. In text-based IR there is no equivalent of a confidence score for individual words — rather, all words deterministically either occur or do not occur. The challenge for SCR is to determine the appropriate way to allow uncertain evidence regarding the presence of the word in an item to contribute to a useful representation of the spoken content item.

Further, during the process of generating the ASR transcript, the ASR system generates information about alternate words that provide good matches for the speech signal. Alternate ASR output represents a rich source of information that can be exploited to improve SCR. Special care must be taken to deal appropriately with the high level of uncertainty associated with the ASR lattices that are used to encode a range of alternate recognizer hypotheses.

The challenge of covering all possible words. The vocabulary of a conventional large vocabulary continuous speech recognition (LVSCR) system is finite, and this vocabulary limits the terms that appear in the speech transcripts and thus the indexing terms that can be used for retrieval. Words that do not occur in the vocabulary of the recognizer are called Out Of Vocabulary (OOV) words. The OOV problem is a perennial challenge for SCR. A larger vocabulary for the ASR system is not the solution, since language is in constant growth and new words enter the vocabulary steadily. In [197], an analysis of news text is presented that demonstrates that vocabulary sizes continue to grow as a data set gets larger. In other words, it is not possible to create a single large vocabulary that will eliminate the OOV problem. Further, under certain conditions, adding more words can compromise the recognition performance of words already in the vocabulary. As pointed out by [174, 197], the OOV problem is particularly disruptive for SCR since many of the new vocabulary words are proper names, which are important for IR. According to [174], up to 10% of all query words can be OOV words, in a typical application that uses a word-based recognizer with a large vocabulary (i.e., LVCSR). Of course it is possible to update the vocabulary of the ASR system by adding new words to the language model. However, as noted by [174], it can be difficult to obtain enough training data to train the language model for new words. Additionally, for most application scenarios, it is not feasible to re-recognize spoken content once the initial transcripts have been generated, due to the high computation costs of the ASR process and the huge sizes of current-day spoken content collections. For these reasons, the OOV problem is a formidable one.

The challenge of context. Speech media is not fully represented by its transcript. Other aspects of the media will impact whether or not a certain result is relevant to the user information need. In the case of video, it is clear that the user may also require the visual channel to fulfill certain specifications. However, even with respect to the audio-channel only, certain aspects might make a particular result more or less suited to the user's needs. The user might require a certain speaker, a certain speaking style, a result of a certain length, speech media of a certain quality (no background noise) or format. Licensing conditions, as with all intellectual property, might also be important. The timeliness of media is important, for example in a news search, the user might want only recent results, or also might want to look back at historical developments. This challenge is also related to how the results are depicted within the user interface. Many of these aspects have bearing on whether or not a user will find a particular result suitable and thus must be represented in the surrogate. Some ambiguities can be handled by the system by simply offering users results of multiple categories in the results list, for example, for the query "Guillermo del Toro," the results might be clips from his films, from interviews and of information about his life.

The challenge of structuring spoken content. Although some speech media is produced in a manner that lends itself to segmentation (think of a news broadcast composed of individual reports), much of it does not have a well-defined inherent structure. It is important not to assume that there exists a predefined notion of a "spoken document" for an SCR system. Thus, the system must either carry out a topical segmentation or it must determine the temporal boundaries of the result that will be returned to the user at query time. The structure is important, since within the retrieval system it is necessary to know which indexing features should be taken into account for the calculation of the ranking score. Also, the results need to be presented to the user in an accessible form.

The challenge of visualization and playback. One of the appealing aspects of full text search involves user assessment of results. A key feature of a successful SCR system is that the user can quickly review speech media results and decide whether or not they are worthy of further attention. In the case of a Boolean search, the result is required to contain the terms that appear in the query. If the user hears a query word very quickly upon initiating playback of a result, it serves as a shorthand confirmation of the relevance of the result. Users are seldom willing to listen to long stretches of audio in order to confirm result relevance. For this reason, intelligent players that allow users to directly confirm the presence of their query words are helpful. Notice that such confirmation is not only important for the purposes of user result assessment, but also to build user confidence in the system. Intelligent players can also contain information about speaker segmentation — such information allows users to gain an impression of the properties and content of the entire result without listening to it from end to end.

The remainder of this survey is devoted to discussing techniques and technologies that allow SCR systems to address these challenges. Along the way, we will refer back to this list of challenges in order to make clear which approaches are helpful for addressing the specific challenges.

3

Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is the technology that transcribes a speech signal into a textual representation. It is also called Speech-To-Text (STT) technology, especially in contexts such as dialogue systems. The designation STT emphasizes that ASR is essentially the inverse of Text-To-Speech (TTS), which is also known as speech synthesis. ASR systems range from isolated word recognition systems used, for example, in command-and-control applications, to Large Vocabulary Continuous Speech Recognition (LVCSR) systems that transcribe human speech in unconstrained human-to-human communication. “Large Vocabulary” speech recognition aims to provide significant coverage of the large and diverse range of word forms used by humans. “Continuous” speech recognition recognizes words in the natural stream of language, where they are generally unseparated by pauses or other cues that could signal a word boundary.

Currently, it is rarely ambiguous whether “speech recognition” means human recognition of speech or automatic recognition of speech. We retain the designation *Automatic Speech Recognition* not only due to the currency of the acronym ASR, but also as a reminder that human beings are the original and the best speech recognizers. As good as ASR systems are or may become in the future, human transcriptionists will remain a highly viable method of generating speech transcripts or spoken content annotations. Ultimately, any fully automatic speech recognition system must compete in speed, robustness, and efficiency with humans or with computer-assisted human annotators.

The material that follows provides background information on the nature of speech and then describes how the speech recognition task is standardly approached, namely, using a technique for probabilistic modeling of time series data called the Hidden Markov Model (HMM) framework.

Many resources exist that can provide a deeper or broader understanding of ASR that goes beyond the basic material presented here. The sections on ASR in Jurafsky and Martin's popular textbook on speech and language [139] are a useful entry point into the field. Other indispensable books in the area of ASR include [90, 117, 124]. The classic reference for HMMs, in the context of speech recognition, is Rabiner's tutorial [222]. Rabiner and Juang's book [221], however, covers additional useful topics. A more recent treatment of HMMs for speech recognition that provides an in-depth discussion of implementation issues is [81]. The present exposition is intended to provide a basic overview that is sufficient for understanding the discussion of ASR in the rest of the survey.

3.1 The Nature of Human Speech

We often produce speech with little conscious thought, but, in fact, human speech is a complex signal with an intricate structure. In order to grasp the complex nature of speech, it is helpful to think of the speech stream as decomposable into progressively smaller units. Real-world speech cannot be cleanly split into elements belonging to such neat hierarchical categories, but this conceptualization is elegant and captures the basic principles. Additionally, it turns out that representations of speech as layers of structure provide the basis for robust and effective speech recognition architectures.

The largest unit of speech is the *utterance*, a string of words produced in a single speaking burst. For example:

> "This union may never be perfect, but generation after
> generation has shown that it can always be perfected."[1]

Although this example happens to be a well-formed sentence, an utterance need not be well-formed or even a sentence at all. Unlike written language, spoken language is produced as a string of sentences only in a very limited number of formal domains. Human-to-human communication is dynamic and speakers change direction mid-sentence. Utterances can also be sentence fragments, a speaker turn in a dialogue, or a run-on. The following is an example of an utterance taken from the AMI Corpus [34], a dataset of transcribed meeting recordings used for multimodal research.

> "There there are time stamps *um* for, well, segments,
> *um* and for th... *um* segments is for example
> when when you look at the data, what is displayed in
> one line."

This example is more typical of an utterance produced during conversational speech.

Utterances are composed of *words*. The conversational speech example above shows that utterances may also contain vocal non-lexical events such as those the speaker produces when hesitating. As a rule of thumb, a speaking rate of more than 240 words per minute is perceived as fast and one of less that 160 words per second as slow [163]. Words are the prototypical bearers of meaning in the language and are the smallest units that can be assigned to lexical categories, also referred to as Part Of Speech (POS) categories. The basic inventory of lexical categories is generally agreed to include eight members: Noun, Verb, Adjective, Adverb, Pronoun, Preposition, Conjunction and Interjection. Inventories that are larger or have slightly different composition are also used, cf. e.g., [7, 139]. The phrase is the basic unit of syntactic

[1] Barack Obama's Speech on Race, March 18, 2009

structure and consists either of a single word, or more typically a group of words. Phrases belong to syntactic categories such as Noun Phrase (NP) and Verb Phrase (VP).

Language processing devotes a great deal of attention to *Named Entities* (NEs). NEs are words or phrases that designate a restricted set of referents. In the most common case, this list is defined to include the names of persons, locations, and organizations. NEs are "named" in the sense that they are proper nouns or noun phrases conventionally used to designate a specific entity, for example, "Knut" rather than "a polar bear." Unlike NPs, which are syntactically defined, NEs are motivated by exigency. The list of NEs can be modified or extended according to the needs of the application [139].

A distinction is often drawn between *function words* and *meaning-bearing words*. Function words are words that play a role in the syntax of the language, but do not have inherent meaning, i.e., it is difficult to provide a classical dictionary-type definition for these words. Function words include articles (e.g., "a" and "the"), prepositions (e.g., "on" and "at") pronouns (e.g., "they" and "she"), and conjunctions (e.g., "and" and "but"). Function words tend to be short, often monosyllabic. They are among the most frequently occurring words, with "the" being the single most frequent word in the English language [139]. Function words tend to belong to the *closed lexical classes*, classes that do not readily develop new members during the natural course of language evolution. Stopword lists used in IR generally target function words. They may also include a variety of other forms that are relatively meaningless due to their frequency in a particular collection.

Meaning-bearing words are classically nouns and verbs. If it is easy to provide a dictionary-type definition or a synonym for a word, chances are it should be considered to be meaning-bearing. Languages demonstrate high levels of linguistic productivity for meaning-bearing words. New words that are coined or introduced from other languages enter into a language in one of the *open lexical classes*, for example, Noun, Verb, Adjective, cf. [110]. Open classes acquire new members during language growth via processes of borrowing or coinage. Vocabulary growth is a basic characteristic of a healthy language. The reality of language change is one of the underpinnings of modern linguistics, cf. the explicit assumption of [67], “Every language, and every dialect within a language, is always in a state of change” (p. 2).

Below the word level, units that cannot stand alone and, for this reason, are generally never used in isolation, can be distinguished. The *morpheme*, like the word, is a meaning-bearing unit. It is the smallest meaning-bearing unit of a language. An example of a morpheme is “dis” in *disfluency*, which conveys the meaning of “not.”

The *syllable* is the smallest unit of speech that can be pronounced independently. It consists of a core nucleus and, optionally, an onset and/or coda. The nuclei of syllables, typically vowels, are usually salient as energy peaks in the speech signal. Nuclei are responsible for the characteristic energy and pitch contours of human speech, which are usually 150–250 ms long [90]. A more detailed discussion of syllable rate is included in [163]. Words and morphemes are units defined with respect to meaning — when we look at these units from the perspective of form, they may turn out to be single syllables such as the morpheme “dis” or the word “speech.”

Prosody refers to variations that do not impact the identity of words, but create meaning or mood nonetheless. Prosody manifests itself acoustically as pitch, lengthening, and loudness. An example is the rising pitch contour that marks a question in English and many other languages. Note that in tonal languages pitch is not prosodic, but rather serves to differentiate the meanings of words.

The *phoneme*, an individual speech sound, is the basic acoustic building block of speech. Phonemes are represented by phonetic alphabets. The International Phonetic Alphabet (IPA)[2] is used by phoneticians. Speech and language systems generally use some variant of IPA, for example, SAMPA (Speech Assessment Methods Phonetic Alphabet).[3] Other examples include the TIMIT phone set cf. [90] and for American English, the ARPAbet cf. [139].

Linguistics distinguishes between *phonemes* and *phones.* A phoneme is an abstract sound category: change a phoneme in a word and the word is either destroyed or has changed its identity. A phone

[2] http://www.langsci.ucl.ac.uk/ipa

[3] http://www.phon.ucl.ac.uk/home/sampa

is a speech sound as it is instantiated in a specific speech signal. When a speech recognition system is designed, a compromise is made between modeling phonemes (i.e., abstract sound categories) and modeling phones (i.e., specific sounds). This compromise serves to explain why speech recognition scientists often use the designation “phone” and “phoneme” interchangeably, despite their distinct linguistic definitions.

Phones are produced when a fundamental frequency generated by an airstream interrupted at quick regular intervals by the vocal cords resonates in the cavities of the vocal tract. The relative size and shape of these cavities are changed by the motion of the articulators (e.g., tongue and lips), giving a speech sound its distinguishing qualities. When the airflow does not encounter any significant obstruction while moving through the vocal tract, a vowel sound is produced. Vowels differ with respect to the configuration of the tongue, jaw and lips, which control the size and shape of the resonating chambers in the vocal tract, for example, the mouth cavity. When the airflow undergoes significant blockage, either in the form of major constriction or complete interruption, a consonant is produced. For example, the phoneme [b] is produced by blocking the nasal cavity and by interrupting the flow of air through the mouth using the lips, cf. e.g., [51, 117, 139] for a more detailed description of phoneme production.

The speech signal changes over time, with the midpoints of phonemes, particular vowels, being the most stable stretches. The transition from one phoneme to the next corresponds to the movement of the lips, tongue and other articulators, which can be visualized as a gesture aimed at achieving a target configuration sufficiently characteristic of a given phoneme to allow humans to distinguish it from contrasting sounds. The fact that articulatory gestures are fundamentally freeform in nature means that it is not possible to define a time point at which one phoneme ends and the next begins. The natural blurring of phonemes one into the other is the source of co-articulation effects in speech. The enormous variability during sound transitions gives rise to a significant challenge for recognizing human speech.

A visualization of a spoken utterance helps to illustrate the magnitude of this challenge. Figure 3.1 is a spectrogram of the speech

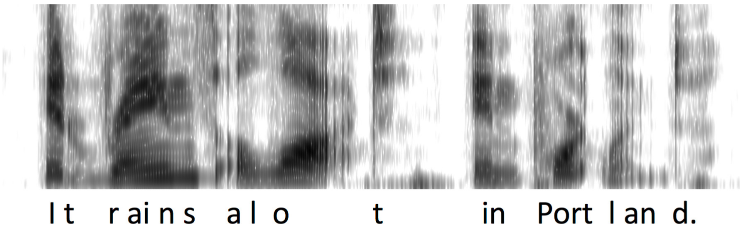


Fig. 3.1 “It rains a lot in Portland.”: Spectrogram of the speech signal.

signal of a human speaker pronouncing the sentence “It rains a lot in Portland.”[4]

A spectrogram shows the variation of the speech signal with time. The vertical axis represents frequency, and the strength of particular frequencies are represented by the relative darkness of their depiction. The horizontal access represents time. It can be seen that the phonemes are not distinct from each other, but rather flow from one into the other.

A single phoneme can be pronounced in different ways, meaning that a single phoneme is associated with multiple phones. These phones are called *variants* or *allophones.* Which allophone is used to realize a particular phoneme is dependent on the neighboring phones or some other feature of its context [51].

3.1.1 The Scope and Variability of Human Speech

In addition to its rich structure, human speech is characterized by a high level of variability. Physiological diversity, individual style differences, speaker origin and the heterogenous and dynamic nature of what we as humans wish to express, all conspire to make the speech signal highly variable and speech recognition a difficult task. A useful inventory of the sources of variability in speech is presented by [20]. This work identifies gender and age as sources of variability due to physiological variation. Social diversity gives rise to non-native accents, which tend to be highly dependent on the individual speaker and also affected by native tongue and proficiency level. It is also the source of regional accents, which tend to be stable over speaker populations.

[4] Credit: Aaron Parecki (http://aaronparecki.com)

The social origins of diversity in human speech style make it clear why human speech is inherently and necessarily highly variable. A speech variety can be defined as the type of speech used by a particular speech community or for a particular function. Sociolinguists define a speech community as a group that uses the same variety of a language and that agrees on appropriate language usage in their social context.

A classic typology of speech varieties is given by [7], which identifies four types of varieties in speech, *standard speech*, *social speech varieties*, *regional speech varieties* and *registers*. *Standard speech* is a variety that is accepted as socially outranking other varieties and is conventionally used in government, broadcast communication media, and educational settings. Standard speech is highly conventionalized, applying rigid norms for pronunciation and syntax, and broadly overlaps with written language. Some language communities establish organizations charged with the definition and enforcement of standard speech, a prime example being the *Académie Française*, the official authority for the French language. Although prescriptivist control of standard speech without doubt has a large impact on speech production, much speech is produced "in the wild" so to speak, compliant with usage convention and dynamic social norms rather than a set of fixed rules. *Social speech varieties* include types of speech used in groups that share a socio-economic status, a gender, an ethnic background, an age background, an occupation, or any other set of common interests or circumstances. Social speech evolves with the dual function of enabling communication between members of the group and differentiating the group from other groups. *Regional speech varieties* arise under the constraints of geographical proximity and *speech registers* are varieties of speech specific to certain contexts or functions. Further information on language variation and change can be found in [36].

In the speech signal, variability in speech can be observed to follow certain patterns. For example, articulation differences among words show a dependency on word identity. Function words are generally reduced and content words are generally stressed, although these tendencies can be reversed in spontaneous speech [31]. However, variability remains very difficult to predict effectively. A fast speaking rate can result from multiple strategies, either speeding articulation or deleting phonemes, depending on style and context [291]. The effect of mood change on speech production is deemed "considerable" by [20]. Results of experiments reported by [291] suggest that the impact of speaking style on speech recognition performance is great.

ASR research distinguishes among different types of speech and defines different tasks. Over the course of the development of the field, tasks of increasing difficulty have been formulated. The earliest ASR systems recognized isolated words only. Isolated word recognition is applied in command-and-control applications such as controlling a car stereo. Research on recognition of continuous speech, that is, words spoken one after another, began in the 1990s. Efforts in the early 1990s focused on recordings of speakers reading text aloud, so-called "read speech." The National Institute of Standards and Technology (NIST) sponsors a series of benchmark tests that have been central to the development of speech recognition technology.[5] Read speech corpora provide good reference material for quantitative evaluation, but do not involve spontaneous effects and also lack the variety in recording conditions (i.e., background noise) characterizing real-world situations [214].

Speech as we deploy it in daily use, which is often *spontaneous speech*, has radically different characteristics from read speech. Upon first consideration, these differences may not be obvious, and in [54] it is noted that speakers are often surprised when presented with a literal transcription of what they have spoken. Particularly striking is the fact that interruptions and corrections are a natural part of conversation. Spontaneous speech effects include variable articulation, variable sentence lengths, disfluencies such as filler words, false-starts, self corrections, and breathing [31, 54]. The description of spontaneous speech in [54] includes unconventional usage of words, agreement mismatches, run-on sentences, hesitations, and restarts that give rise to partial words and phrases. Non-verbal communication such as eye contact and nodding behavior is also considered by [54] as part of spontaneous speech — although this information is not available to a speech recognition system that only processes audio input.

[5] http://nist.gov/itl/iad/mig/

Recently, progress has been achieved in a number of particular domains of ASR including telephone speech [96], children's speech [220], noisy environments [58], speech emotion recognition [244] and meeting speech [23]. In the next subsection, we turn to the details of how an ASR system is built.

3.2 The Hidden Markov Model Framework

ASR is the process of transcribing an acoustic signal into text. In its most basic form, this text is the sequence of words $W = w_1, w_2, w_3, \ldots, w_n$ comprising the spoken content of the signal. The fundamental task of the recognizer is to determine which word sequence W best matches the sequence of acoustic observations $O = o_1, o_2, o_3, \ldots, o_n$. The HMM approach was first applied to speech recognition in the 1970s [16, 223]. In the 1980s, statistical approaches to the task of speech recognition came into their own and the HMM approach emerged as the dominant approach to LVCSR [138]. These systems were groundbreaking in the area of recognizing words spoken in a natural speaking flow by an arbitrary speaker. HMMs provide a unified statistical modeling framework that exploits the structure of speech and at the same time captures its variability and temporal dynamics. The system processes a speech signal with the goal of making an optimal decision concerning the sequence of spoken words. Probabilistically, speech recognition is expressed as finding the word sequence $\hat{W}$ that is most probable given the sequence of acoustic observations O that compose the original signal.

$$\hat{\mathbf{W}} = \arg\max_{\mathit{All}\ \mathbf{W}} P(\mathbf{W}|\mathbf{O}) \tag{3.1}$$

$\hat{W}$ is the recognizer output and is often referred to as the recognition hypothesis. This expression is expanded using Bayes' Law to yield,

$$\hat{\mathbf{W}} = \arg\max_{\mathit{All}\ \mathbf{W}} \frac{p(\mathbf{O}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{O})}. \tag{3.2}$$

$P(\mathbf{W})$ is the prior probability of a sequence of words occurring and is provided by the language model. $p(\mathbf{O}|\mathbf{W})$ is the acoustic likelihood of a string of observations giving a sequence of underlying words and is provided by the acoustic models, which are trained at the phone level and then linked into words using the lexicon. The operation involves comparing the hypothesis word strings only in terms of their relative scores. For this reason, the denominator $P(\mathbf{O})$, which remains the same for any given sequence of acoustic observations, can be dropped when calculating the recognizer output, yielding,

$$\hat{\mathbf{W}} = \arg \max_{All\ \mathbf{W}} p(\mathbf{O}|\mathbf{W})P(\mathbf{W}). \tag{3.3}$$

The elements of Equation 3.3 correspond directly to components of the HMM speech recognition system. Figure 3.2 summarizes in a block diagram the HMM framework for a typical speech recognition system. In the following discussion, each module of this diagram will be treated in turn.

3.2.1 The Language Model

The language model encodes the probability of one word being spoken after another. In [279], the challenge of language modeling is described as finding a balance between maximum constraint of the search space and maximum freedom of the input. The standard language modeling approach for LVCSR is to use the so-called n-gram model. The language model is trained using large amounts of text, and encodes the probability of the occurrence of a given word, based on the words before it. The list of language model innovations that have been used to improve speech recognition efficacy includes higher order n-grams,

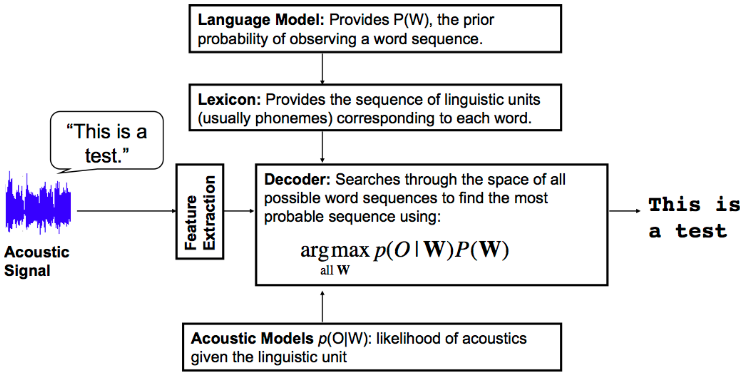

Language Model: Provides P(W), the prior probability of observing a word sequence.
Lexicon: Provides the sequence of linguistic units (usually phonemes) corresponding to each word.
"This is a test."
Acoustic Signal
Feature Extraction
Decoder: Searches through the space of all possible word sequences to find the most probable sequence using:
$\arg\max_{\text{all } \mathbf{W}} p(O \mid \mathbf{W})P(\mathbf{W})$
This is a test
Acoustic Models p(O|W): likelihood of acoustics given the linguistic unit

Fig. 3.2 Block diagram of a typical HMM-based speech recognition system.

for example 5-gram language models in which each word is conditioned on the context of the previous four words, class-based language models and multi-gram models. Language models are smoothed by introducing corrections to the estimates of term occurence frequency. Smoothing is an important issue for language models and we refer the reader to the literature [43] for more detail.

Collateral text can also be used to extend the vocabulary of the speech recognizer and to adapt the language model [13, 224]. In this process, text can be collected from suitable available sources such as news feeds, domain-specific publications or Internet sources. The material is analyzed to identify significant new words, which, if they are deemed important enough, can be added to the vocabulary of the recognizer. In order for a new word to be added, a pronunciation must be generated and a probability must be estimated for the language model. Word pronunciations can either be researched in a suitable dictionary or potentially generated using grapheme to phoneme conversion [272]. Some ASR systems use multiple pronunciations for individual words and pronunciation modeling has long been recognized as a key issue for ASR [261]. Similarly, language model probabilities can be checked in a table of counts from the existing language model training data, if the word was present in the training set but not selected for inclusion in the vocabulary, or otherwise using some estimation method.

Another technique for improving the language model with additional information is the “Web 2.0” approach to speech recognition introduced in [93, 207]. Here, ASR transcripts and speech media are made available in parallel on the Internet and users are encouraged to correct the mistakes in the transcripts. The transcripts are then used to re-train the speech recognizer. Crowdsourcing in this way has great potential to provide hugely expanded training resources for language-based applications at reasonable cost.

3.2.2 The Lexicon

The lexicon encodes the pronunciations of the words in the speech recognizer vocabularies. The pronunciation of a word is represented by a string of phonemes. A single word can be associated with multiple pronunciations. If a word is not in the lexicon, it cannot be recognized by the speech recognizer and is designated as an *Out-Of-Vocabulary* word or OOV word. OOV words can never occur in speech recognizer output, but instead will be substituted by whatever word or word sequences from within the available lexicon that yield the highest probability output during the recognition process.

3.2.3 Acoustic Modeling

Each speech sound is represented by an acoustic model. If a language can be covered by a relatively small inventory of syllables, as is the case for Chinese and Japanese, syllables and not phones are used as the basic unit for acoustic modeling [117]. In other cases, however, acoustic models are trained at the phoneme level. The phone inventory for English speech recognition usually contains around 50 phones [117].

Three rules for choosing acoustic units for speech recognition are outlined in [117]. The unit should be *accurate* (able to represent the sound in its interchangeable contexts), *trainable* (it should be possible to collect sufficient training data), and *generalizable* (the sound should play a part in providing coverage for new words that are introduced into the vocabulary). In practice, if an ASR system is implemented using an existing lexicon, the acoustic model inventory will be limited by the transcription alphabet used by the lexicon. Because the phoneme inventory and the realization of particular phonemes varies from language to language, it is usually necessary to train a new set of acoustic models for each new language that is to be transcribed by the speech recognizer. However, situations exist in which multilingual phone sets are appropriate [147, 198].

A typical speech recognition system represents each phoneme with a set of triphone models. A triphone is a phone together with the context provided by its right and left neighbors.

Triphone inventories are kept compact by “state tying”, which identifies similar neighborhoods to conflate for the purpose of parameter estimation. Acoustic models are trained by using large amounts of acoustic data that have been transcribed. A typical amount of training data would be 300 hours and upwards. An important consideration is

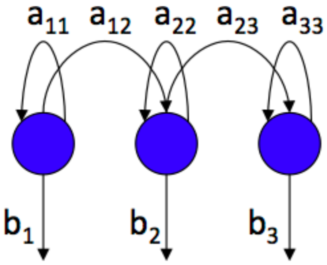


Fig. 3.3 Acoustic model of a phone.

whether the speech data used for training is well matched with the speech data that will be transcribed by the ASR system. For example, training data should reflect the types of speakers and channels conditions anticipated in the data to be recognised.

Some statistics of phone lengths in the Switchboard corpus (a dataset of recorded spontaneous telephone conversations) are given by [139]. The statistics reveal phonemes to be highly variable in length, with short lengths as little as 7 ms and long lengths as much as 1.3 seconds. The basic acoustic model represents phones in a length-independent fashion. As a result, it does a good job of generalizing over phonemes pronounced at different speeds, but may also sacrifice important information encoded in the length as a result.

3.2.4 Feature Extraction

The speech signal is transformed into a sequence of acoustic feature vectors before it is processed by the speech recognition system. The vectors contain spectral features, which encode how much energy is present at different frequencies in the speech signal [139]. These vectors are extracted from overlapping windows of the speech signal. Typically, each second of speech is represented by 100 spectral feature vectors. Each vector is extracted from a signal window that is small enough to support the assumption that the speech signal is stationary (non-changing) in its duration. A typical window is 25 ms in length. For each vector, the window is shifted forward by, for example, 10 ms. This overlap seeks to ensure that rapid changes in the input signal are captured in the feature vectors. In general, acoustic vectors are 39 components in length. Commonly, the components are mel-frequency cepstral coefficients (MFCCs), the change of the MFCCs between windows, and the change of the change of the MFCCs. Mel frequencies are frequency bands warped to approximate the sensitivity of the human ear. Perceptual linear prediction is also commonly used for spectral vectors — this is a linear prediction method that retains information in the signal that is relevant for human perception.

3.2.5 Training

Training is the process of estimating the parameters of the HMM. A corpus of transcribed speech is necessary for training. A special case of Expectation Maximization called the Baum–Welch algorithm is used to estimate emission probabilities of individual states and transition probabilities between states. Emission probabilities are commonly modeled as mixtures of Gaussians. It is possible to bootstrap acoustic models by performing a rough recognition and by using the resulting transcripts to feed model training — cf. the discussion of lightly supervised and unsupervised training in [81]. If speaker-specific data are available or become available, acoustic models can be adapted to speakers [81].

3.2.6 Recognition

Recognition is the process of determining the most likely sequence of words spoken given an observed speech signal. Typically, the "noisy channel model" is used to conceptualize the recognition process. Under this model, the underlying sequence of words spoken is taken to be the original signal and this sequence is considered to have been distorted by noise during the generation of the speech signal. The goal of recognition is to recover this original signal, and, for this reason, recognition is often referred to as "decoding." The recognition process involves searching through the large space of word strings that could be possibly hypothesized by the speech recognizer in order to find that string that best matches the input. This process is referred to in the speech recognition community as "search" and the output of the speech recognizer as the "hypothesis." In the area of speech retrieval, the word "search" is generally reserved for the process of retrieving results in response to user queries and, less frequently, for the process of speech recognition.

Decoding strategies fall into two categories, time-synchronous and time-asynchronous. The Viterbi algorithm is the main example of time-synchronous decoding. Time-synchronous decoding applies a breadth-first strategy for exploring the search space, considering each hypothesis possible at a given point of time before moving on to the next point of time. Time-synchronous methods make use of dynamic programming algorithms, which reduce the computational complexity of a solution of the overall problem by breaking it in to incrementally solvable sub-problems. Time-asynchronous decoding pursues a depth-first strategy, in which the best hypotheses are explored forward in time before being compared to their competitors. The distinction between time-synchronous and time-asynchronous decoding is not absolute [12]. In practice, time-synchronous decoders limit the number of hypotheses under examination at any given point using a process called beam pruning, which eliminates unpromising hypotheses from consideration. Asynchronous stack decoding attempts to expand the best hypotheses first rather than pursuing a purely depth-first strategy.

In the next subsection, we turn to discuss some aspects of ASR systems that are particularly relevant when ASR is used in order to realize SCR.

3.3 Aspects of ASR Critical for SCR

In this subsection, we discuss forms of recognizer output beyond the 1-best transcript, the use of subword units for ASR and also the nature of error in ASR transcripts. These topics provide more specific background material for the discussion of "Considerations for the Combination of ASR and IR" in subsection 3.4 and also the discussion of the exploitation of ASR output in *Exploiting Automatic Speech Recognition Output.*

3.3.1 Recognizer Output beyond the 1-Best Transcript

The output produced by a speech recognizer can be represented in different forms. The choice of forms is dependent on what the output will be used for. The simplest form is called the first best hypothesis (1-best transcript) and consists of a single string of words, for example, `using` `text factors for search`. This string represents the word sequence identified by the recognizer as being the most likely one pronounced in the input speech signal. The recognizer can also be set so that it also outputs hypotheses that are less likely than the first best, but nonetheless likely. Such output is called an n-best list. A typical value for N is 10 and an example of a 10-best list is illustrated in Figure 3.4.

A lattice, pictured in Figure 3.5, is a network that encodes likely hypotheses of the recognizer. The horizontal positions of the nodes

```
using text factors for search
using text vectors for search
using text vectors for surge
using tags vectors for surge
using tags factors for surge
use syntax vectors for search
use syntax factors for search
use syntax factors for surge
using text fact or surge
using tags fact or surge
```

Fig. 3.4 Illustration of an n-best list with $n = 10$.

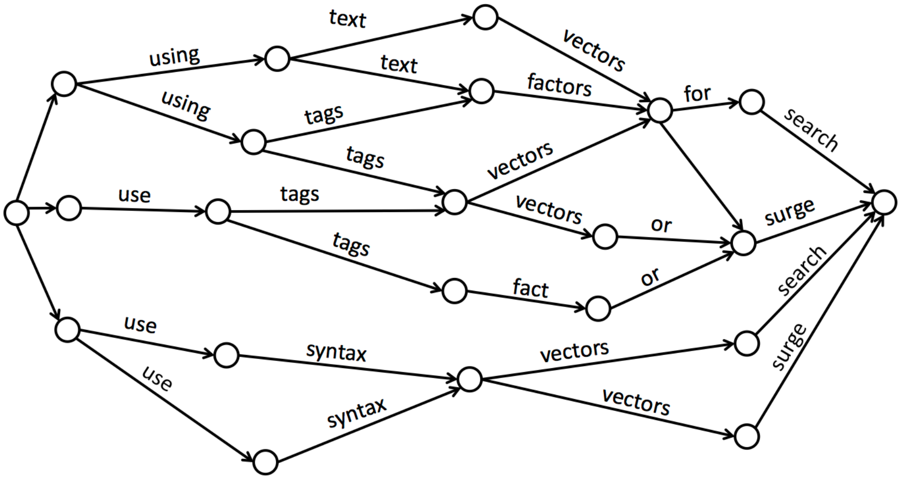


Fig. 3.5 Illustration of a word lattice. Hypotheses with different start and end points are represented as different nodes.

encode time information. Each node corresponds to the start time of a hypothesized spoken unit. In the general case, the recognized units are words, but they can be any base unit used by the language model of the ASR system, for example, syllables or phones. The distinguishing feature of the lattice is that it preserves the time information generated by the speech recognizer. In other words, if the top-n list contains the same word form recognized with two different starting times, the word is represented as two nodes in the lattice. Each possible path through the lattice corresponds to a recognizer hypothesis. The lattice makes it possible to store a large number of speech recognizer hypotheses using a reduced amount of memory.

A confusion network, pictured in Figure 3.6, is a lossy representation of a lattice. It is constructed such that it contains a path for every path that existed in the original lattice, but it discards time information generated by the recognizer. Instead, it retains only information about the relative position of recognized units and their competitors. It is not necessarily the case that all the alternatives in the set compete on exactly the same time segment within the speech signal. However, the word confusion network (WCN) is formed so that the confusion sets of words are close enough to be regarded as competing and the confusion network is normalized so that the probabilities of all competing words sum to one [81].

The structuring of words into a series of sets gives the confusion network a distinct form, which is often thought of as reminiscent of a string of sausage links. WCNs are usually created with an algorithm, originally proposed by [177], that bears the nickname “sausages.” Indexing lattices and WCNs for SCR is discussed in greater depth in *Exploiting Automatic Speech Recognition Output.*

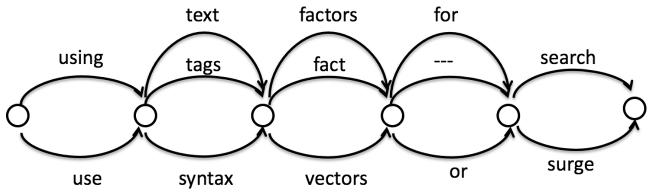


Fig. 3.6 Illustration of a word confusion network.

Speech recognition transcripts can be improved by combining the output of multiple recognizers. A prime example is *Recognizer Output Voting Error Reduction* or ROVER. Combination techniques such as ROVER [73] are dependent on the fact that different recognizers produce different kinds of errors. Dynamic programming alignments create a word transition network and a rescoring or "voting" process is used. Within the resulting word transition network there are correspondence sets. Voting between the alternatives in the correspondence set is carried out on the basis of frequency of occurrence, average confidence scores and maximum confidence scores.

3.3.2 Subword Recognition

In the case of an LVCSR system, the vocabulary of the recognizer consists of lexical words. Other sorts of ASR systems, however, can make use of base units other than lexical words. Two challenges must be faced in order to use a base unit other than an orthographic word. First, the training text for the language model must be decomposed into the alternate units so that the n-gram probabilities can be estimated. Second, appropriate pronunciations must be assigned to the alternate units.

A particularly attractive choice of base unit is a morpheme, the smallest unit of meaning in the language. A morpheme-based recognizer for the English language would, for example, contain the units "develop," "test," "ing," "ed" and "ment" and would be able to combine these units to recognize the forms "developing," "testing," "developed," "tested" and "development." The advantage is that a small number of units covers a larger word-level vocabulary. The disadvantage is that the language model bears a much larger burden of preventing illicit combinations from being recognized, for example, "testment." Also, the pronunciation of morphemes is sometimes context dependent (cf. the pronunciation of "ed" in "developed" vs. in "tested"). For this reason, it is sometimes advantageous to retain the association with the base words.

There are two basic strategies for splitting words. First, they can be split according to statistical occurrences of phone strings, an approach chosen by [296], which refers to such strings as particles. Second, they can be split according to linguistic rules.

Two types of linguistic splitting can be distinguished. The first is compound decomposition, that is, decomposing “keyboard” into its component pieces “key” and “board.” Compound splitting results in word fragments (i.e., compound components) that can be considered to be content bearing, in the sense that they can be assigned a meaning. In English, compounds generally split into stand-alone words, as is the case with the example of “keyboard.” In other languages, such as German and Dutch, compounding is a highly productive form of word generation, meaning that new compounds are created freely “on the fly” rather than being subject to particular rules or conventions. In such languages, compound components are often special combining forms that cannot be used as stand-alone words. For example, the German compound *Spracherkennung*, “speech recognition,” decomposes into *sprach*, “speech,” and *erkennung*, “recognition.” The stand-alone form of “speech,” however, is “Sprache” and not “sprach.”

The second kind of linguistic splitting breaks up a base form and additional morphemes. Words can also be split into a base form plus derivational morphology (e.g., “utterance” = “utter” + “ance”), in which case the base form has a different part of speech than the original word. In other cases, they are split into a base form plus inflectional morphology (“recognizes” = “recognize” + “s”), in which case the part of speech does not change. In some languages, such as English, there is a strict limit on how many morphemes can be added to a base form. Other languages are agglutinating, meaning that long strings of morphemes can be added to base forms. Examples are Turkish and Finnish.

The choice of base unit for the language model is often determined by the use to which the ASR transcripts will be put afterwards. Using smaller units in the language model makes it possible to provide better coverage for the input signal — the recognizer will encounter less speech content that it cannot cover with its language model. The disadvantage of using smaller units is that the system gives up much of the constraint on the search space. When the search space can contain only words, which are relatively long, possible transitions from unit to unit take place less frequently. Adding compound components to the recognizer vocabulary means that the recognizer is capable of producing nonwords, which could be disturbing to a human reader, but yet may be useful in the context of a speech indexing application. We return to discuss extracting indexing features from subword ASR transcripts in *Exploiting Automatic Speech Recognition Output* (subsection 4.4).

3.3.3 ASR Error

ASR output is usually evaluated in terms of Word Error Rate (WER). WER is calculated by calculating the number of word-level mistakes in the speech transcripts. Word-level mistakes fall into three categories: insertions, substitutions, and deletions. Examples, again taken from the AMI corpus of meeting recordings [34], illustrate each type of error. An insertion error occurs when the ASR system hypothesizes a word for which there is no corresponding word spoken in the speech signal. Here, the recognizer has inserted the word "so":

Spoken:	"That's	at	the	end.		That's..."
Recognized:	`That's`	`that's`	`the`	`end`	`so`	`That's`

A substitution error occurs when the word hypothesized by the recognizer does not match the one spoken in the speech signal. In the previous example, "at" was substituted by the recognizer with "that's."

The following example illustrates a deletion error:

Spoken: "to all those who need it"
Recognized: `To all those need it`

Here, the recognizer has dropped the word "who," which does not appear in the transcript.

In order to calculate word error rate, the hypothesis string generated by the ASR system is compared to the reference string containing the words actually spoken in the speech signal, using the minimum edit distance. Word error rate is a combination of insertions, substitutions, and deletions.

$$\mathbf{WER} = \mathbf{100} \times \frac{\textbf{Insertions} + \textbf{Substitutions} + \textbf{Deletions}}{\textbf{Total Correct Words in Transcript}} \tag{3.4}$$

A dynamic programming method is used to create an alignment of the reference transcripts with the speech recognizer output.

Errors can be roughly divided into two categories: errors that occur because a word spoken in the speech signal is not present in the recognizer vocabulary (so-called OOV errors) and errors that occur due to a mismatch between the acoustics produced by the speakers and those expected by the recognizer. The exact source of error cannot be explicitly determined, but human listeners can usually make a good attribution to a particular category. For example, in the error analysis in [31] the following categories are used: the acoustic model, the language model, the articulation quality of the segment and other effects such as breaths, out-of-vocabuary words or extraneous noise. The authors of [31] conclude that non-disfluent spontaneous speech has the same error rate as read speech. Here, disfluencies include pause-fillers, word fragments, overly lengthened or overly stressed function words, self-edits, mispronunciations and overly long pauses. In [31] it is recommended that systems explicitly model these events in the recognizer's lexicon.

These examples illustrate disfluencies, which can give rise to ASR errors. The first contains a self-edit and the second contains a filler word, which is correctly recognized, but associated with an error in its vicinity.

Spoken: "everything that's a word has a st... time stamp."
Recognized: `everything let's to work as as tight time stamp`

Spoken: "There there are time stamps *um* for, well, segments"
Recognized: `That there's times sense um for low segments`

Note that the word "stamp" is recognized correctly in the first utterance but that "stamps" is mis-recognized as "sense" in the second. As discussed in detail later, words that are repeated in the speech signal help to compensate for recognition errors because the ASR system has multiple "opportunities" to get them right. If these two utterances were contained in the same spoken content item, and if stemming were applied to pre-process the transcript for indexing, the confusion between "stamps" and "senses" would lower the count of the stem `stamp` in the index, but the item would still be able to match a query that contained the word "stamp."

It cannot be expected that the length of the words or the number of syllables will match up in the spoken utterance and the recognized utterances. Most implementations of HMM recognition do not model length explicitly and, as such, spoken words can be contracted during substitution error.

Spoken: "be on utterance level"
Recognized: `Beyond the front so`

Again, item length is an issue. If items are long enough, the number of words in the transcripts provides a good approximation of the number of words spoken in the speech signal. These examples of ASR error provide a lead-in for the next subsection, which begins with a discussion of the impact of ASR error on SCR performance.

3.4 Considerations for the Combination of ASR and IR

The issue of how to combine ASR and IR optimally has occupied researchers quite intensively. We argue that consideration of the combination of ASR and IR will allow an SCR system to outperform one in which ASR and IR have been combined naïvely. In this subsection, we first look at the implications that ASR error has for SCR. Then we turn to discussion of design decisions during implementation which are important for the integration of ASR and IR.

3.4.1 Impact of ASR Errors on SCR

The relationship between the quality of the transcripts produced by an ASR system and the performance of a spoken content retrieval (SCR) system that makes use of these transcripts is a complex one. Conventionally, transcript quality is measured in terms of Word Error Rate (WER), defined in subsection 3.3.3. The WER is dependent on the domain. In [221], a textbook published in 1993, the state-of-the-art in ASR performance was reported for fluent speech recognition on the Naval Resource Management corpus. The WER was on the order of 4%, which is roughly the same level of error made by humans carrying out manual transcriptions of naturally produced speech. This bright picture comes nowhere near the reality of the speech data in the domains for which SCR systems are currently developed. The vocabulary of the Naval Resource Management corpus was slightly less than 1000 words and the recording conditions were highly controlled. As a result, this low error level is not typical for real world applications.

Interestingly, and possibly unexpectedly, experience has shown that a SCR system can make use of transcripts with a relatively high WER without compromising retrieval performance. The robustness of SCR to ASR error was studied in the literature as early as 1994 [240]. In [104], experiments were carried out with simulated WERs where it was shown that WER of up to 50% still provided 85%–90% SCR accuracy. In [5], effectiveness of retrieval is reported to fall less than 10% even with WER of around 40%.

The large corpus of broadcast news used for the TREC Spoken Document Retrieval track in 1999 and 2000 (i.e., SDR at TREC-8 and TREC-9) was recognized with WERs that ranged from 10% to 20%. In [38], it is noted that TREC SDR shows that retrieval performance was not impacted by these error levels, but that WERs of 30%–50% are typical for more challenging domains. In [148], WERs of 10%–20% are cited for spoken news and of 20%–40% for voicemail and conversational speech. In [39], a WER of 50% is reported for state-of-the-art recognizers on recorded lectures used for SCR. WERs of ca. 50% are faced by [176], in the case of spoken document retrieval from call-center recordings, and by [47], in the case of retrieval from conversational telephone speech. In short, high ASR error rates are characteristic of many of the domains for which SCR systems are developed and systems that assume very high quality speech transcripts may be of limited applicability in real world SCR settings.

3.4.2 Suitability of WER for Predicting SCR Performance

In practice, WER may not be the best measure of quality of speech transcripts and their usefulness in a retrieval setting. In [256], it is stated that WER is not particularly appropriate for use in the retrieval context because it treats all words equally. In the area of spoken language understanding, it is shown that language models that produce better understanding accuracy actually suffer from *higher* WERs [284].

In [83], it is found that the correlation between WER and SCR performance was non-linear for the 1998 (TREC-7) TREC SDR data collection. The literature contains several proposals for what is referred to by [82] as "IR-customized ASR scoring," in other words, metrics that capture those aspects of ASR transcripts that make them well-suited for SCR. In [83], several alternatives were explored. Filtering stopwords from the ASR transcripts before calculating WER did not succeed in achieving a better correlation between WER and SCR performance. The WER calculated on the basis of Named Entities (people, locations, and organizations), however, did show a better correlation with SCR performance than WER alone.

3.4.3 Interaction of ASR Error and IR

Although standard WER might not be the best predictor of SCR performance, what is clear is that ASR error does impact SCR. In [237], it is shown that ASR transcripts that provide good matches with queries tend to have lower WERs than ASR transcripts that match the query less well. Analysis showed that documents that contained a broader range of query words tended to have lower WERs. The authors conclude that it is the process of retrieval itself that is identifying documents with lower WERs. This is not a surprising result. If a topic can be considered to be characterized by a certain pattern of word use, then it can be expected that the pattern will characterize both a human generated query related to that topic and also human speech on that topic. The pattern is better represented if the ASR WER is low and the query can be more easily matched with the content item.

The importance of topical word patterns, leads us to expect that errors generated by ASR have a good chance of having minimal impact on SCR. WER patterns are influenced by factors (such as acoustics) that are independent of the human process of meaning production. Because of this independence, the distribution of word errors will not readily fall into characteristic patterns for topics. In other words, it is not obvious that mis-recognized words could easily appear in a pattern that "accidentally" resembles a topic. It is an open question whether, and under which conditions, the distribution of word errors in ASR transcripts will mimic topic patterns. It is reported in [5] that automatically expanding TREC SDR queries using ASR transcripts seems to add helpful recognition errors that are correlated with query topics — these helpful words are effectively representing in the query a spoken word that the ASR system consistently misrecognizes. In [162], an analysis of turn-sized segments of meeting speech from the AMI corpus is carried out. The results suggest that "habitual" errors of the ASR system (when a word is consistently substituted for another word) have less of an effect on the overall semantic relatedness between items in the collection than "infrequent" errors, i.e., substitution errors the ASR system makes no more than three times in the entire collection.

The literature analyzing the statistics of mis-recogized words in ASR transcripts is restricted to relatively few references. In [240], it is noted that errors produced by the recognizer and well-recognized query features have different distributions and that IR models are able to exploit this difference. In [135], human-generated transcripts and ASR transcripts from the 1999 TREC SDR (TREC-8) were compared with respect to word statistics. It was found that the human transcripts contained roughly four times more unique word types than the ASR transcripts. The ASR transcripts contained a radically smaller number of singularities (word types that only occur once in the collection). The authors point out that misspellings in the human transcripts can impact this number, but the effect remains marked. The picture that emerges is that the impact of WER on the distribution of word types in ASR transcripts involves a large number of types that are deleted or substituted out and a rather smaller number of "favored" types from the ASR system's vocabulary that are inserted or substituted in.

Word errors potentially have unexpected interaction with IR models. Recall from the overview of IR models in subsection 2.2 that methods for calculating term weights and term probabilities typically take account of term frequencies and document lengths. These term weights are frequently integrated into IR models using non-linear functions. The implication is that the first occurrence of a word in a document can be very important, but subsequent occurrences may make only a very marginal difference in the ranking function. Word types that are infrequent and are substituted out by the speech recognizer, possibly because they are not in the vocabulary, pose a clearer threat to SCR performance. Words missing in the ASR transcript destroy the possibility of creating a match between queries and items.

For longer queries and documents, the impact on ranking is likely to be small due to redundancy effects. Essentially, longer queries and documents provide richer expressions of topical word patterns and more reliable matches. In [5], the challenge of short queries and short documents is pointed out. The mismatch issue is likely to be most significant for short queries consisting of only two or three words. If even one of these is absent from the transcript then it can have a significant impact on the matching score, reducing the relative score of the item and its rank. Similarly, a single match with an error in the transcript could cause an irrelevant item to be ranked high in the list. This effect is similar to what is observed for standard text IR in cases in which there is mismatch between query terms and terms used in the document — although in the case of SCR noise in the form of errors, the transcript will probably cause an even greater negative effect than in the text case.

In sum, ASR error leads to reduction in retrieval effectiveness in SCR. The underlying reasons for this arise from a complex interaction of query matching issues and distortion of term weights. These issues can to some extent be overcome by a range of techniques, including external text collections to provide better collection frequency estimates, and document and query expansion techniques to ameliorate matching problems. These techniques will be discussed further in *Spoken Content Retrieval beyond ASR Transcripts.*

3.4.4 Strategies for Combining ASR and IR

Implementing an SCR system requires making a decision concerning the level of integration between the ASR component (that transcribes the speech signal to text) and the indexing component (that derives useful indexing terms from the text) used for IR. In the following discussion, we briefly sketch the arguments for isolating and then for integrating ASR and indexing in an SCR system.

Isolating ASR and IR. A system that isolates ASR and retrieval, does not allow information concerning the retrieval process to inform the generation of the speech recognition transcripts. The primary motivation for isolating ASR from retrieval in an SCR system is modularity. The speech recognizer can be treated as a black box: it produces a certain output and this output is then passed to the indexer. The indexer extracts terms from the transcripts and creates the index that will be used for search.

Maintaining modularity greatly simplifies the implementation and maintenance of an SCR system. It enables, for example, the use of an off-the-shelf ASR system, which is easy to use, but may offer only a limited amount of control over the language model, the acoustic models and the system parameters. When ASR and indexing are fully isolated, it is possible to substitute one ASR system for another with a minimum of effort. This flexibility could be an advantage if new and better ASR systems or systems versions become available.

Isolation has another practical advantage. If the ASR system is isolated from the indexing process, it is possible to optimize it just once, and from that point leave it to do its job, representing lower maintenance costs. LVCSR is an incredibly complex process. In [12], a "trilogy" of factors is identified as being responsible for the best-performing LVCSR systems: combination of multiple algorithms, clever design cooperative with the hardware, careful parameter tuning. If any of these factors goes awry during the process of integrating the ASR and indexing system, sub-optimal performance could result.

It can be argued that there exists a precedent for maintaining modularity. The success of the HMM framework for ASR is built upon a high level of separation between individual components. The acoustic models are separated from the language model and both can be trained and optimized separately. They are brought together (via the pronunciation lexicon) during the transcription process. The dependency of phoneme pronunciation on word identity or sentence structure is modeled by using context-dependent triphones, but otherwise neglected. Historically, this separation has proved efficient and productive for speech recognition, suggesting that judicious modularization is also the appropriate tactic to pursue in SCR systems.

Finally, it is important to consider the needs of the user interface. The interface might require transcript information, in addition to the terms that are used for indexing. For example, full transcripts or snippets might be needed for display to the user. In this case, the ASR system needs to generate a classic transcript, closely resembling written text. If the ASR system and the indexing system are integrated, it is necessary to ensure that the proper text material is also generated for representing spoken results in the interface.

Integrating ASR and IR. The primary motivation for integrating ASR and indexing in an SCR system is information exploitation. Additional information made available during the transcription process serves to restrict the search space of the speech recognizer, leading to a better recognition result. This principle has made an important contribution to the success of speech recognition in the area of spoken dialogue systems. In general, dialogue systems do not carry out LVCSR, but rather use a highly specified model of what a human can possibly say during that section of the dialogue, making heavy use of wordspotting and grammars. The success of dialogue systems can be used to support the argument that SCR systems should strive to identify and extract the semantics of spoken content during the ASR process, rather than waiting for the indexing process.

Proponents of a high level of integration between ASR and indexing remind us that speech recognition has a critical dependency on speech understanding. Indeed, as humans we are well acquainted with the difficulty of following a conversation on a topic that we know nothing about. Alex Hauptman has pointed out that speech recognition cannot be perfect without understanding and that understanding is considered to be "AI-complete," namely, as difficult to solve as the problem of artificial intelligence itself [102]. If a system uses natural language processing for another task within the video library, it should be tightly coupled with the speech recognition component. The exigencies of system design often prevent tight integration of ASR and information retrieval within a system.

ASR is a resource intensive process and the extreme case is clear: the collection must be transcribed in advance — ASR cannot run at query time. However, approaches that generate subword lattices using a first recognition pass and subsequently re-score these lattices as additional information becomes available are computationally viable approaches. These are important approaches to be aware of and will be treated in more detail in *Exploiting Automatic Speech Recognition Output*. Such approaches make it possible to address *The challenge of covering all possible words* caused largely by OOV words, which are not in the lexicon of the ASR system and therefore can never be recognized. They also address *The challenge of handling uncertainty* — the variability of speech means that there is never a perfect match between the speech stream and the ASR systems models.

Design of an SCR system that integrates ASR and IR requires a solid knowledge of both components. If a linguistic resource already used during the ASR process is to be used again during the IR process, it must be used in a different manner. If not, the same information will have simply been added to the system twice and performance will not improve. Information should be added to the system where it can be most helpful. If a resource used by the IR component could be more effectively exploited by the ASR component, systems should be designed to use it there. In particular, the trend towards producing lattice representations and also to integrating long range syntactic dependencies into the decoding process, as mentioned by [12], is an important one.

Historically, ASR research has pursued the vision of "the listening typewriter," a machine that would produce a transcript comparable to that of a human transcriptionist using a typewriter. Spoken content, however, is not structured in the same way as text and does not lend itself to representation in a conventional transcript, unless it was dictated for that purpose. Although ASR error rates have been pushed to dramatically low levels for speaker-dependent applications, it is only user groups with specialized needs who rely on them heavily for transcription. These groups include doctors, who transcribe large volumes of information with specialized vocabulary and conventionalized structure, and users, who cannot manipulate manual interfaces or who need to save hands and wrists from undue strain. In only a small number of cases, does ASR actually serve the function of "the listening typewriter."

Despite the limited success of this initial vision, LVCSR shows promise of yet living up to its potential as a revolutionary technology. SCR has become an important application area for LVCSR. In 2003, Ken Church predicted that search as a form of consumption would replace dictation as a speech and language processing priority [50]. If ASR research is increasingly pursued with the goal of SCR, a tighter integration between ASR and indexing will become a natural development in the future.

Achieving an optimal balance. In sum, a system that isolates ASR from indexing should strive to carry out the best, most flexible ASR transcription when the spoken content is initially processed. The ASR transcripts should retain as much information as possible (e.g., lattices, which encode alternative recognizer hypotheses) so that, as further information becomes available, it is still possible to make use of detailed information from the speech signal. The transcription should be suitable both for indexing purposes and for the generation of representations of documents and results in the user interface. The optimal balance between ASR and IR will depend on the use scenario in which the system is to be deployed. If, for example, the speech is recorded under ideal conditions and OOV is not a problem, then it may be the case that the first-best word-level output of an LVCSR system may be entirely suitable to provide users with the speech content that they are looking for and further optimization aimed towards integration of ASR and IR is not necessary. This example is exaggerated since most domains in which SCR systems are needed are far from being highly ideal. However, it serves to illustrate the danger of seeking a naïve or a "one size fits all" solution for the integration of ASR and IR in an SCR system.

We now turn to examine specific techniques and technologies that can be drawn upon for developing SCR systems. Specifically, we next treat in greater detail the issue of extracting indexing features from spoken content and using these features for SCR in a way that achieves an optimal balance between ASR and IR.

4

Exploiting Automatic Speech Recognition Output

Here, we cover techniques for exploiting the output of an Automatic Speech Recognition (ASR) system in a spoken content retrieval (SCR) system. We focus on the indexing of ASR transcripts. Indexing is the process of creating representations of content items that capture their key characteristics and make it possible to access them using a search system. The representation of an item consists of a set of indexing features, which in the case of spoken content includes words or word-like units, and, importantly for SCR, information on where these units appear (i.e., their order or time codes). As in text-based IR, such indexing features are called "terms." A term can be considered a string of characters used to represent a content item. IR systems do not treat all terms equally, rather terms are associated with different weights or probabilities in the system. SCR makes use of corollaries of term counts and other term-level statistics that play a central role in classic text IR frameworks (cf. subsection 2.2). In the SCR setting, calculation of these quantities additionally can involve taking into account the information contained in the ASR output that is able to reflect uncertainty concerning the correspondence of this output to the original spoken input (i.e., the words spoken in the original speech signal).

The indexing techniques presented in the material that follows make it possible to adapt and apply text IR indexing approaches to ASR transcripts. As a first step to understanding the indexing challenges in the SCR setting, it is helpful to reflect on the appropriateness of the word "transcript" to designate the output of an ASR system. "Transcript" implies a transcription of spoken content that respects the conventions of written language, that is, it is a human-readable text. Although ASR output is, in many cases, human readable (i.e., consists of lexical words), in a large number of cases it is not. In subsection 3.3.1, we saw that ASR output can differ from conventional text, in that it is characterized by noise (both word errors and segment boundary errors) and it contains timing information. Differences between conventional text and ASR output correspond to additional information sources that could and should be exploited for SCR. We focus our discussion on three sources of information often included in ASR output and their use in SCR:

- the inclusion of *multiple recognizer hypotheses*
- the association of each unit with a *confidence score* that represents its level of match to the speech signal
- the use of *subwords* as the basic unit of recognition

In subsection 3.4, we arrived at the conclusion that it was necessary to strive for a balance between ASR and IR that is specific for a particular SCR system and its use scenario. These three information sources represent three opportunities for achieving an effective balance of integration between the output of the ASR system and the index used for SCR. In the following discussion, the usage of these information sources is covered and their untapped potential is highlighted.

This material addresses the first two challenges for SCR that were listed in subsection 2.4. We begin by discussing techniques that have been developed in response to the first challenge, *The challenge of handling uncertainty.* Subsection 4.1 "Going beyond the 1-Best ASR Output," deals with techniques for indexing n-best lists, and lattices and confusion networks, structures that encode multiple recognizer hypotheses. Subsection 4.2 "Exploiting Confidence Scores," presents the use of information on the reliability of the ASR output for indexing.

The remaining subsections discuss techniques that are concerned with representing uncertainty, but which also address the second challenge, namely, *The challenge of covering all possible words*. Subsection 4.3, “Representing Speech with Subwords,” gives a general introduction to sub-lexical units that can be used for vocabulary independent representation of spoken content. Subsection 4.4, “Subwords as Indexing Features,” covers methods of extracting subwords from speech, presenting a survey of techniques that use subwords and combinations of subwords as the basic indexing unit for SCR. Subsection 4.5, “Hybrid Approaches for SCR,” describes how subword-based and word-based approaches can be combined for improving SCR performance.

In *Overview of Spoken Content Indexing and Retrieval*, a distinction was made between SCR, which focuses on meaning-based relevance ranking, and other “searching speech” tasks that aim to find mentions of query terms, for example, Spoken Term Detection (STD). Approaches that find mentions are, however, in the SCR setting for the purpose of extracting indexing features. The final subsection 4.6, “Techniques for Detecting Spoken Terms,” provides a brief overview covering these approaches.

The techniques presented next can be used to compensate for the limitations and also to exploit the potential of the information produced by the ASR system during the speech recognition process, thus potentially improving SCR performance. Other sources of indexing features that can be used for SCR, such as metadata and social annotations, will be examined in *Spoken Content Retrieval beyond ASR Transcripts*.

4.1 Going beyond 1-Best ASR Output

As described in subsection 3.3.1, ASR is capable of producing several forms of recognition output. The recognition process involves determining the 1-best hypothesis, that is, the most likely word string spoken in the audio signal. During this process the ASR system generates information about “runners-up,” other highly competitive hypotheses. Information about the most likely hypotheses can be output in the form of a list (n-best list), but also in the form of a lattice or a confusion network. The motivation to use information derived from hypotheses beyond the 1-best output of the ASR system is straightforward: words spoken in the speech signal not contained in the top-one most likely hypothesis might indeed be contained in other highly likely hypotheses. In [247], it is noted that for a given utterance, there are generally a small number of top hypotheses produced by the speech recognizer that are close in likelihood. The 1-best hypothesis is, of course, the most likely of these, but is typically not strikingly more likely that the other top hypotheses, pointing to the potential high value of the "runners-up."

In [39], a set of lecture recordings is investigated for which the ASR system achieves a 1-best Word Error Rate (WER) of 45%. The authors calculate the WER of the lowest WER path through the lattices generated by the ASR system to be 26%. They note that this ratio might not generalize to other systems or content, but conclude that it suggests that lattices do have enormous potential to improve SCR recall. The use of lattices is advocated by [238], which notes that using the 1-best hypothesis is probably sufficient for tasks with moderate error rates (about 20%), but that tasks with higher error rates require multiple ASR hypotheses. Whether or not multiple hypotheses can yield performance increases will also be dependent on the domain and on the task, for example, whether it involves retrieval of short speech segments or longer documents.

In this subsection, we survey methods that have been applied to exploit multiple hypotheses of the ASR system for SCR, looking in turn at n-best lists, lattices and confusion networks. The subsection covers only those methods that use word-level indexing features and exact matching techniques. Discussion of methods that make use of multiple hypotheses consisting of subwords and of fuzzy matching is delayed until after we have treated the use of subwords as indexing units in subsection 4.4. Note also that the techniques discussed in this subsection are used for meaning-oriented SCR. Techniques for mention-oriented search, such as STD, also make use of lattices, and these will be discussed in subsection 4.6.

4.1.1 Indexing n-Best Lists

Using n-best lists is a simple approach to exploiting multiple recognizer hypotheses that has been reported to work well in the literature. Work by Siegler et al. at CMU [245] showed that using the top 50 hypotheses of the recognizer dramatically improved the performance of their spoken document retrieval in the 1997 TREC-6 SDR task, which involved a corpus of English-language broadcast news containing 1500 stories (ca. 50 hours of spoken data) and a set of 50 queries for known-items (i.e., only one document in the collection was considered relevant to the query). The 1-best WER of the ASR system was approximately 35%. Inverse Average Inverse Rank (IAIR) of the topmost relevant documents in the lists returned by their Vector Space Model (VSM) retrieval system improved from 1.37 to 1.32. Siegler's dissertation [247] reported on experiments on the 1997 TREC-6 task in which extremely long n-best lists were used. The task involved 23 queries and a corpus containing ca. 2800 documents (ca. 75 hours of spoken data). The use of long n-best lists yielded SCR performance that was improved 64% over that achieved when only the 1-best hypothesis was used. The approach assumes that all hypotheses in the n-best list are equally likely, but also imposes the assumption that the N hypotheses are independent of each other and the individual words in the hypotheses are also independent. Weights for the VSM are calculated simply by concatenating the top-n hypotheses for each spoken document and counting the terms according to a standard text-IR procedure. The approach is an appealing one, since it provides a very simple estimate of term frequencies, obviating the need for the more complex models needed to derive term frequencies from lattices or confusion matrixes. These experiments were carried out on a relatively small corpus of broadcast news pre-segmented into stories. The gains of this method may or may not transfer to the larger, less well-behaved corpora currently used for SCR research. However, the simplicity of using n-best lists makes it worthwhile to investigate, or at least consider, this approach in any new SCR use scenario.

4.1.2 Indexing Lattices

Lattices provide a compact representation of top-hypotheses of the ASR system. Recall from *Automatic Speech Recognition* (cf. Figure 3.5) that a lattice is a directed acyclic graph structure that stores a speech recognizer search space that has been pruned to contain the most likely output hypotheses. The first lattices used in SCR were actually phone-lattices, and their use dates back to the earliest SCR systems developed in the mid-1990s [120, 132]. We will return to discuss these techniques in subsection 4.6 after we have also covered subword units such as phones.

The lattice contains the most likely hypotheses of the speech recognizer and in this way resembles an n-best list. Unlike the n-best list, however, it contains timing information — in other words, the time point at which the recognizer hypothesized each word is encoded in the lattice. In [37, 38, 39], a 17%–26% improvement in Mean Average Precision (MAP) is reported compared with indexing the 1-best recognizer output. The corpus used contains 169 lectures of approximately one hour each. The 1-best WER of the ASR transcripts was ca. 45%. The authors proposed a "soft hits" approach that explicitly addresses the fact that existing schemes for indexing lattices are not designed to take into account information concerning the proximity of words.

They point out that there are two aspects of uncertainty about a word in a lattice: first, whether or not that word was actually spoken and, second, the position at which it appeared (including its ordering relative to other words). On individual paths through a lattice, position indexes can be assigned to words in the order in which they occur. The uncertainty arises in calculating which position should be considered the overall position of the word within the lattice for the purposes of applying information retrieval (IR) approaches that make use of word proximity. Proximity information is needed to search for phrases, for example, or to integrate nearness constraints into multiple word queries.

The "soft hits" approach [37, 38, 39] involves reducing the lattice generated by the ASR system into a "Position Specific Posterior Lattice" (PSPL), which is smaller than the original lattice and also makes possible indexing of word position. The challenge of deriving word-position information from a lattice can be best understood by noting that on any given path through the lattice, the order of the words on that path is uniquely defined. However, given one word in the lattice, it is not possible to assign it an overall position, since it may have different positions on different paths. The PSPL lattice bins the words in the original lattice and stores the probability that a word occurs in a particular position. A *soft hit* consists of an integer position in the speech segment and a posterior probability. The score of a word is calculated by summing its probability across all positions. These scores are then used as weights in a VSM retrieval scheme. A major advance of PSPL is that existing implementations of phrase matching can be applied to the PSPL index without modification.

An alternative to PSPL called "Time-based Merging for Indexing (TMI)" is proposed by [315]. Recall that PSPL retains information about the relative order of words, but discards timing information. The TMI approach is designed with an eye to applications requiring navigation and previewing, in which the time stamp of individual words is important. Like PSPL, TMI lattices achieve dramatic improvement in MAP over approaches using 1-best. On the same lecture recording data set described above, 1-best achieves a MAP of ca. 0.53 and TMI of 0.62. TMI reduces the lattice by merging word hypotheses that have nearly identical start and end times. TMI outperforms PSPL in terms of the trade-off between the index size and accuracy.

There are two specific characteristics of the lecture recording set used to evaluate PSPL and TMI that should be noted. First, the speech style of the speakers falls between planned and spontaneous, meaning that it is a challenging corpus, but that a completely spontaneous corpus would prove more challenging. Second, the soft hits were calculated on the basis of sentence-length segments that were produced by aligning the speech signal with a transcript containing human generated segment boundaries. The performance of the approach under less constrained conditions remains an empirical question.

In [47], a lattice-based approach is proposed and tested using the BM25 model and the language modeling framework. This work includes a careful analysis of the impact of lattice pruning on lattice-based SCR. The optimal pruning setting is a trade off between insertions and deletions of the correct word in the lattice. In order to achieve good SCR results, it is important to carefully select the degree to which the ASR output lattices are pruned before indexing. The methods were tested on a corpus of English-language telephone speech containing nearly 2000 hours of data divided into approximately 12000 conversation-sized documents of up to 10 minutes in length.

The 1-best WER was ca. 50%. The highest performance was achieved using lattices: with lattices, a MAP of about 0.77 was reported and could be achieved using both the BM25 model and the language modeling framework. The added advantage of using lattices is quite slim, however. Applied to 1-best transcripts, the language modeling approach already achieves a MAP of 0.76. These results suggest that the advantages of using lattices are dependent on the use scenario, including the task, the data set and the ASR system.

4.1.3 Indexing Word Confusion Networks

Word Confusion Networks (WCNs) are simplifications of lattices that allow direct access to information about competing word hypotheses. Recall from *Automatic Speech Recognition* (Figure 3.6) that the nodes in a confusion network do not correspond to points in time, as the nodes in a lattice do. Instead, they correspond to starting and ending points of words and their hypothesized alternatives. Each arc in a WCN is associated with the posterior probability of a word. Each path in the original lattice has a corresponding path in the WCN. Approaches that index WCNs for SCR build on [177], which proposes an algorithm for deriving confusion networks from lattices. The WCN is, in essence, a multiple alignment of all the different paths contained in the lattice. The first step in creating a WCN is initializing the equivalence classes. The initialization is done by creating classes from all words that are the same and have the same starting and ending times. The clustering algorithm then merges equivalence classes, taking identity and phonetic similarity into account. The posterior word probabilities within the WCN are then scaled and normalized to sum to one, over all hypotheses. For an alternate approach that aligns all other paths to the 1-best path through the lattice refer to [98]. The WCN is very similar to the TMI structure discussed in the previous subsection. A practical difference is that WCNs can contain null transitions [177], which makes them a less natural choice for phrase matching (i.e., matching of adjacent terms).

Previous to their application to SCR, WCNs were shown to yield performance improvement in the area of spoken language understanding [97]. In [176], WCNs were applied to SCR and experiments were conducted using a large corpus of English-language call center data. Experiments were carried out using the VSM framework and results were reported for WER levels in the range of 30%–65%. The data set contained around 2000 calls of approximately 10 minutes each. Results were evaluated by comparing the retrieval performance on the ASR transcripts with retrieval performance on corresponding human-generated transcripts. These experiments demonstrate a relationship between higher WER and deteriorating SCR results. Using WCNs decreases precision over the use of 1-best transcripts, but overall MAP is increased when weights derived from WCN are boosted with information on their relative rank within their equivalence class. Further exploration of this approach, including analysis of the impact of pruning on the weights used for WCN indexing, would help to shed light on whether, in practice, the use of WCNs is worth the computation involved to extract them.

In sum, the relative merits of using word lattices have strong dependencies on the domain, the ASR system, lattice pruning, and the way in which lattices are processed for indexing. It remains difficult to build a system that takes advantage of multiple recognizer hypotheses that is better-principled and outperforms that achieved with n-best lists. Intuitively, lattices and WCNs make it possible to incorporate information about word hypotheses that are in direct competition with each other. However, it was pointed out in [247] that a word is contained in many good paths through the lattice, then its direct competitors will be contained in relatively fewer good paths in the lattice. In this way, scores calculated on top-n lists do capture to some extent the competition between different words in the same part of the speech signal. Any indexing of lattices or WCNs must incorporate a substantially better manner of encoding the uncertainty of individual words if they are to lead to overall SCR improvement.

An important aspect of the approach proposed in [176] involves taking into account the confidence level of each word. Because of the potential for confidence in support of SCR, we turn in the next subsection to a brief overview of the area of confidence score computation for speech recognition transcripts.

4.2 Confidence Scores

A confidence measure is a score that encodes the ASR system's assessment of the reliability of its output [125]. Word-level confidence scores reflect whether a word actually occurred in the spoken content where it was recognized. The scores express the reliability of the hypotheses or their probability of correctness. In a provocatively-titled 1996 paper, "Towards increasing speech recognition error rates" [22], a case is made for "knowing what we do not know." This paper called for more effort to be invested in methods for estimating when the speech recognizer goes wrong, rather than in just making errors less frequent. This call is motivated by the observation that the success of ASR technology is in task completion. In the SCR setting, task completion involves providing a set of items that satisfies the user's information needs. Fifteen years later, we have at our disposal an array of interesting techniques for estimating word-level confidence in ASR output, but no single widely-used framework that guides the integration of this information into the SCR system in such a way that guarantees improvement. This subsection provides a compact overview of methods for generating confidence scores, in order to provide an introduction to the topic for researchers who are interested in using these scores to improve the performance of spoken content retrieval (SCR). We also note the existence of helpful surveys on confidence scores, including [125, 294].

In general, confidence scores fall into two categories, depending on the kind of information used to calculate them: scores using information from the ASR system and scores integrating information from external sources. In the discussion that follows, we describe each method in turn, and cover techniques that have been used for incorporating confidence scores into SCR systems.

4.2.1 Information from the Speech Recognizer

Probabilities derived from the ASR system are a widely-used source of confidence information. Recall from subsection 3.2 that the speech recognition process involves searching for the most likely path through a search space encoding all possible word strings known to the recognizer. The string that is output by the recognizer (i.e., the 1-best hypothesis) is more likely than any other hypothesis, but the recognizer does not calculate its likelihood in absolute terms. Early work on ASR confidence scores [311] characterized the information derived from the ASR system with the remark, "We know which utterance is most likely, but we don't really know how good of a match it is."

The process of looking for the relatively most likely string corresponds to the simplified version of Bayes' rule in which the denominator is dropped (cf. Equation 3.2 vs. Equation 3.3). During the recognition process, different word-string hypotheses are compared with respect to a given segment of the speech signal, that is, with respect to the same acoustic observations. For this reason, it makes sense for an ASR system to calculate only the relative likelihood of the hypotheses — the normalization introduced by the denominator is not necessary. The situation changes, however, when we want to generate a confidence score. A confidence score should reflect the overall probability of a word having been correctly recognized and not ignore dependencies on characteristics of the speech signal. The un-normalized scores generated by the ASR system are not suited for comparing word hypotheses that were generated by the recognizer as fitting different acoustics [294]. Instead, we need a normalized score, in other words a true posterior probability of the recognized words. The normalizing factor is the prior probability of the acoustic signal $P(O)$, whose calculation is made tractable by applying a method for approximation.

Methods of generating confidence scores from speech recognizer output generally differ with respect to the approximation that they choose. In [311], normalization is accomplished by carrying out a second recognition of the speech signal using a phone-based recognizer and using the score derived in this way as an approximation for $P(O)$. Such approaches are motivated by the following reasoning: $P(O)$ can be calculated by summing over $P(O|W)$ for every possible hypothesis W. This sum can be approximated by determining $P(O|W)$ for the most closely competing hypotheses, in this case those of the second phone-based recognizer.

Other researchers have focused on approaches in which only one ASR system is necessary. A basic form of normalization can be accomplished using the n-best list. In [290], a word score is generated by summing the likelihood of all hypotheses containing the word and dividing by the total likelihood of all hypotheses. A related approach, used by [294], makes use of a posterior word probability calculated on a word lattice. Here, the probability for a word in the lattice, i.e., a single lattice arc, is calculated by combining a sum of all the probabilities of all possible histories of the arc and all possible futures of the arc, normalized with respect to the total probability mass in the lattice. Note that if the language model probability and the acoustic probabilities are all set to one in the lattice, then this approach reduces to counting the number of paths through the lattice that pass through the word, normalized by the total number of paths through the lattice [294].

One of the challenges of lattice-based scores involves making a decision about which words will be considered to compete within the lattice. A word with few competitors should have a higher confidence than a word with many competitors. In the work of [142], a simple density-based approach is applied. The approach examines the speech signal spanned by each word in the 1-best ASR output. At each frame within the word, the number of competing links within the lattice is counted. This count is equivalent to the number of links that intersect a vertical line drawn through the lattice at a time point corresponding to a particular frame. Scores are calculated using various combinations of these statistics.

Another basic challenge is determining which hypothesized words should be treated as a realization of the same underlying spoken word [290]. In essence, the use of lattices in confidence score generation requires overcoming the same issues confronting the use of lattices for indexing, addressed in the previous subsection. In [294], word probability scores are accumulated for words in the lattice that are determined to correspond to the same word spoken in the signal.

The issue of confidence scores is addressed both in the literature on Large-Vocabulary Continuous Speech Recognition (LVCSR) and in the literature on STD. Recently [280, 281], a discriminative approach to estimating the confidence of words recognized by a STD system has emerged. The discriminative approach is motivated by the conjecture that acoustic and language models perform poorly at modeling OOV terms and that generative approaches to confidence are sub-optimal in these cases. The strong performance delivered by the discriminative approach on OOV terms supports this conjecture.

4.2.2 Information from Other Sources

Information from sources external to the recognizer can be integrated for the purposes of confidence estimation. We provide a survey of the nature and the diversity of these information sources by overviewing selected work of authors who have contributed in this area. Early research was carried out within the context of detecting mis-recognitions in dialogue systems. In [311], semantic, pragmatic and discourse-based information sources are combined with acoustic information to build a confidence measure for this purpose. In [68], word-level acoustic scores were combined with context-independent and context-dependent features. The context-independent features included the number of phones per word, the frequency of word occurrence in the language model and acoustic training data sets, and statistics on the occurrence of the phones in the word in the acoustic training set. The context-dependent features included the length-normalized likelihood of the sentence according to the language model, the estimated signal-to-noise ratio, and the speaking rate.

An extensive inventory of information sources for the purposes of estimating confidence was evaluated by [239]. It includes features already mentioned, as well as scores reflecting additional characteristics, such as language model backofff during decoding, the number of active final word states in the ASR system search space, the number of pronunciation variants and the frame-wise average entropy of the phone-level acoustic models. In [302], a computationally cheap score derived from the language model is proposed and combined with acoustic measures. Finally, in [179], information about language model backoff is combined with acoustic information and applied not only for the purpose of detecting well-recognized words, but also of well-recognized segments.

4.2.3 Integrating Confidence Measures into IR Models

SCR makes use of confidence scores by integrating them into IR models, as introduced in subsection 2.2. We now turn to discuss the approaches that have been taken to achieve this integration. It should be noted that since confidence scores are often derived from information contained in lattices and confusion networks, it is not possible to draw a sharp line between techniques that integrate confidence measures into IR models and techniques that index lattices and confusion networks (i.e., techniques already discussed in subsection 4.1), and some overlap in issues is to be expected.

There are two basic categories of method that have been applied to integrate confidence scores into IR frameworks. In the first, an attempt is made to change the counts of terms that provide the input for the statistics (weights and probabilities) for the IR model. This approach often amounts to a computation of "effective frequencies" that then serve as a replacement of term frequency and document frequency counts. In the second, a hard threshold is applied and only words with confidence scores above this threshold are considered to occur in the spoken content. This approach then uses standard word statistics, as applied in text IR, to calculate counts. In general, both approaches require tuning using additional labeled data, in order to optimize performance.

The use of expected occurrences for the purposes of SCR has its roots in early tasks involving filtering speech messages. At the beginning of the 1990s, an early system [230, 231] classified speech messages according to topic class. Keyword likelihoods, weighted by learned sigmoid functions, were used as input to the message classifier. In [181], an "expected number of occurrences" of keywords in messages is obtained by summing the scores of keyword hypotheses that are generated by a word spotter. Thresholding is a common practice for keyword spotting systems and also for word features derived from lattice scans as in [122, 132]. The particular setting of the threshold is important for maximizing SCR performance, cf. e.g., [77, 131]. In [288, 289], a collection of spoken documents is transcribed using a recognizer that generates phoneme sequences and "query features" (words in the query) are extracted from these transcripts at query time. Individual occurrences of query features are called slots. A hard threshold is imposed on slots, by ranking all the detected slots in the collection by their probabilities and discarding all by the top-n most probable. This procedure amounts to a hard cutoff threshold. The use of keyword spotting and STD to extract features for SCR will be touched on again in subsection 4.6.

The choice of whether to impose a hard threshold or whether to integrate a confidence score into the retrieval model depends on the use scenario for the system and the domain of application. If the use scenario requires a very high precision list, admitting items containing words recognized with low confidence might hurt system performance and a tight, hard cut at a high threshold might be helpful. If recall is important, such items might help to ensure that as many relevant documents as possible are retrieved from the collection. Additionally, if the domain of application is structured such that the units of retrieval are short sentences rather than longer spoken documents, redundant use of words within one document will be limited. It is then not possible to rely on there being other words in the document that would compensate for 1-best errors. In this case, it might be useful not to impose a high threshold. In the discussion that follows, we overview techniques that have been used to integrate confidence scores into IR models for SCR, especially those based on acoustic information. The picture that emerges is that exploitation of confidence scores is difficult and that there is probably not one overall solution that is suitable for every situation.

In [245], an oracle experiment is performed, which assumes access to perfect confidence information, that is, direct knowledge about whether a word was correctly or incorrectly recognized. All words in the speech transcripts not occurring in the reference transcripts are removed. This leads to a small improvement in SCR performance: the Inverse Average Inverse Rank (IAIR) of the topmost relevant documents in the lists returned by a VSM retrieval system decreases from 1.38 to 1.33. This experiment suggests that, in general, confidence scores should be helpful. However, the authors are unable to achieve the same performance when automatic estimates of confidence are used. Other contemporary authors also report difficulties in effective exploitation of confidence scores. In [236], the authors report failure of acoustic information, reflecting the value of recognition confidence of words to improve retrieval, but conclude that their approach has run afoul of a length normalization issue.

A more formal analysis of the incorporation of confidence measures in term weights is described in [128]. This work examines the impact on *idf* and *tf* functions for the VSM and BM25 that arises from replacing binary 0/1 presence indicators with normalized confidence measures. The analysis suggests that the impact of confidence measures on term weights will generally be very small, and often outweighed by the effect of speech recognition errors. Modifications of the standard *idf* and *tf* functions were derived and evaluated for the BM25 weighting. The results showed that term weighting using only the modified *idf* function, i.e., a standard binary independence weight, produced a small improvement in retrieval effectiveness; however, this improvement disappeared when it was integrated into the full BM25 model. Experiments showed the *tf* function to have a detrimental effect on retrieval. These results are generally consistent with other attempts to incorporate confidence measures in term weighting for SCR.

The literature also contains instances of successful attempts to exploit confidence scores. In [246], word lattices are used to calculate a "Lattice Occupation Density" (LOD) score, similar to the lattice density measure previously mentioned above. Known-Item-Retrieval experiments were carried out. A training set was used to derive a model of word probability from the LOD score. The model generated probabilities for each word in the 1-best hypothesis, which were then used in a VSM with probabilistic weights. The word probability model improved retrieval performance on ASR transcripts measured against retrieval performance on reference text transcripts.

Confidence information derived from WCNs has been exploited for the purposes of retrieval. The authors of [176] report that for retrieval using 1-best transcripts, methods incorporating information on the confidence level outperform term frequencies. The improvement is slim, but does appear to increase for higher levels of error rate. A question that remains open is whether a better optimization of lattice pruning might have led to better performance of the term frequencies approach.

In [208], an approach is proposed for estimating the frequencies of vocabulary-independent terms for spoken content indexing. The method uses a discriminative technique to estimate frequencies for phone sequences and is evaluated by comparison to phone sequence counts derived from reference transcripts. Since this approach is a vocabulary-independent approach, it also provides an appropriate lead-in for the following subsection. We turn from approaches that focus on *The challenge of handling uncertainty* to approaches that also address *The challenge of covering all possible words.*

4.3 Representing Speech with Subwords

The challenge of covering all possible words is generally referred to, more prosaically, as "The OOV Problem." Recall that, in order to recognize a word, an ASR system must have that word included in its vocabulary. For many SCR use scenarios, it is not possible to assume that all necessary words can be known in advance. For this reason, SCR systems suffer under the problem of OOV words — words that are encountered in the speech signal, but are not contained in the vocabulary. In order to address the OOV problem, words are not recognized directly, but rather in smaller building blocks, called subwords. This subsection presents an introduction to subwords that will provide the necessary background to the discussion in the material that follows.

4.3.1 Introduction to Subwords

A subword is a unit that is smaller than an orthographic word. It may or may not correspond to a linguistic unit (cf. subsection 3.1), such as a phoneme, syllable or morpheme, but it may also be any small unit that is used by an SCR system. The main motivation for SCR systems to use subword indexing units is to address the OOV problem. A subword inventory represents the speech stream with a smaller, more flexible set of building blocks, which gives greater coverage of the spoken content than a predefined vocabulary. This subsection provides a background on the principle of subword units.

Subword indexing units also have the potential of addressing the general problem of error. Although OOV is a major contributor to

ASR error, acoustic mismatches and sub-optimal pronunciation modeling can lead to the recognizer mis-recognizing an in-vocabulary word, substituting another word or word string in its place. The substituted word string represents a "best fit" with the signal and for this reason stands to share a large degree of similarity with the correct word. Subwords make it possible for the retrieval system to exploit partial matches (also referred to as "inexact matches" or "fuzzy matches") within speech transcripts. These partial matches are particularly useful in situations where unexpected pronunciations or channel conditions cause a recognition error in the ASR transcripts.

The creation of a subword inventory for the representation of spoken content follows one of two basic strategies: either subwords are based on the orthographic forms of words or on word phonemizations. *Orthographic subwords* are derived from the written forms of words and consist of a sequence of graphemes. For example, under the orthographic subword approach "knowledge" would be represented as `know ledge`. The advantage of this method is that a text corpus can be easily decomposed into subwords for the purposes of training a subword language model for the speech recognizer. Component orthographic subwords are similar for words with similar spellings. In this example, "knowledge" (`know ledge`) shares a common subword with "knowing" (`know ing`) and with "acknowledging" (`ac know ledg ing`). In the case of orthographic subwords, it is necessary to provide their pronunciations to the ASR system in the lexicon. A single subword can have multiple pronunciations. For example, orthographic subword `know` is pronounced differently in "knowledge" and "knowing."

Phonetic subwords are derived from word pronunciations. Here, subwords are represented as strings of phonemes. For example, under the phonetic subword approach "knowledge" would be represented as `n_A_ l_I_dZ_`. In this case, component orthographic subwords are similar for words with similar pronunciations. Under the phonetic subword approach, "knowledge" (`n_A_ l_I_dZ_`) does not share a common subword with "knowing" (`n_o_ w_I_N_`)[1] Notice,

[1] The underscore, '_', in the phonetic representation is used for readability, but also can be used to differentiate word-internal from word-initial and -final forms in the vocabulary.

however, it shares two common subwords with “acknowledging” (`I_k_ n_A_ l_I_dZ_ I_N_`). Under the phonemic subword approach there is only a single pronunciation per subword. The number of different orthographic subwords that map to a phonetic subword with a single pronunciation varies from language to language. Another important language-dependent effect is the number of shared subwords resulting when semantically related words are decomposed.

The error compensation potential of subword units can be best illustrated by considering an example. We choose to examine the orthographic subword decomposition of the word “Shenandoah.” The same principles apply to other words and to phonetic subwords. Two possible subword decompositions of “Shenandoah” are syllables (`she nan do ah`) and overlapping grapheme strings (`shen hena enan nand ando ndoa doah`). Systems that make use of subwords are attempting to leverage two effects.

First, if a word is not in the vocabulary of the ASR system, it can be reconstructed from a series of subword units. For example, the original audio could be recognized using an ASR system with a syllable vocabulary. If this vocabulary contained the units `she nan do ah`, it would be able to recognize the string `shenandoah` without explicit knowledge of the existence of the word “Shenandoah.”

Second, if the ASR system makes an error with a particular word, subwords can help to provide a partial match. Take again the example of the ASR system with a syllable vocabulary. If an error occurs, for instance, because the word is spoken in the speech signal with a pronunciation differing slightly to that included in the ASR system’s lexicon, the following string might result `she nen do ah`. In this case, three out of the four syllables are recognized correctly. The possibility to make use of a partial match during the IR process is left open. If the ASR system had a word-based vocabulary, it would output a word-level error for the misrecognized word, for example, `crescendo`. This word-level error is difficult to match with the original spoken word, “Shenandoah.”

It is also possible to take this partial match one step further. Note that the syllables in our mis-recognized strings `she nen do ah` (output of the syllable-level recognizer) and `cre scen do` (syllabified output of the word-level) contain syllables that are very similar to the correct syllabification of the word `she nan do ah`. Specifically, `nen` is not far from `nan` and `scen` bears a resemblance to `she`. Some subword-based systems attempt to use matches on multiple levels, in order to create the most reliable inexact match possible between the target word and its realization in the ASR transcripts.

It is not necessary to have a recognizer with a subword language model in order to exploit subword matching effects. A word-level transcript containing `crescendo` in place of `shenandoah` could be decomposed. A syllabification (syllable-based decomposition) of the substitution error word would yield the syllable sequence `cre scen do`. Here, one out of four syllables matches a syllable in the original spoken word. This match is rather distant, but could still prove useful to the retrieval system. Using subword units, however they are generated, thus implements a partial match between words.

4.4 Subwords as Indexing Features

The choice of indexing units is critical to the performance of an ASR system. Strong statements of their importance were made early on: in 1995, [241] asserted "... that indexing vocabulary is the main factor determining the retrieval effectiveness, and the recognition performance is a minor factor once it achieves a certain quality" (p. 11) and in 1997, [304] identified OOV as the "... single largest source of retrieval error" (p. 31). In this subsection, we examine indexing features that are extracted from ASR transcripts and used to retrieve speech content. We look first at word-level features and then at subword-level features. In each case, we discuss in detail the potential for these features to address *The challenge of covering all possible words*.

In *Automatic Speech Recognition*, we introduced the linguistic units that are used for speech recognition. In a conventional English-language ASR system, phone-based units (e.g., triphones) are used for acoustic modeling and lexical words are used as base units of the language model. However, subsection 3.3.2 introduced the notion that other choices of linguistic units may be appropriate for ASR systems and highlighted the advantages of using language modeling units that are smaller than words.

The trade-off between the advantages and disadvantages of using subword units is quite similar, whether they are used as the base-units of the language model in ASR or as indexing features in an SCR system. Subwords are short and modular, meaning that due to the structure of language, a finite, or highly limited set of subwords provides the building blocks to generate all possible larger units. For these reasons, subwords make possible a better fit with the spoken content in a speech signal. The high coverage provided by subwords means that a greater proportion of the information occurring in the original speech signal is retained in the ASR transcript. The preserved information creates an opportunity to confront the OOV problem, that is, *The challenge of covering all possible words.*

Features of all shapes and sizes, from phonemes through morphemes to word phrases, have been investigated at some point in the literature, and the most useful cases will be reviewed in the material that follows. However, it is important to keep the disadvantages of subwords firmly in mind. A major disadvantage of subword units is that recognition rates fall as units get shorter, since context is modeled less robustly. Another disadvantage is the ambiguity that they introduce. Ambiguity is a classic problem that must be confronted in text-based IR. Examples are simple English-language examples including double-meaning words like “fly” (insect vs. airborne movement), “mouse” (rodent vs. computer pointing device), and “Cambridge” (UK vs. Massachusetts). If these words are to be considered in terms of subwords, for example, `cam+bridge`, additional confusion is introduced, namely, confusion between other forms containing the same components, for example, “drawbridge,” which decomposes as `draw+bridge`. The success of subword indexing features for improving SCR depends on the ability of the system to exploit the greater coverage while compensating for the additional ambiguity.

Subwords that are appropriate to serve as a set of indexing features for SCR must satisfy several requirements. We formulate a list of requirements that is loosely based on the criteria laid out in [87]:

- *Representative.* The units should be able to adequately represent the speech stream. The inventory should provide adequate coverage.

- *Discriminative.* The units should allow for semantic discrimination of different parts of the speech stream (or between different speech documents, if documents are defined within the collection).
- *Accurate.* The speech recognizer must be able to reliably extract the indexing units from the speech stream.
- *Efficient.* The representation should require a reasonable amount of computation at ASR-time and be very fast at search time.

In fact, this requirements list is not particularly specific to SCR or to subword units. Indexing units for text-based IR should, quite obviously, also be representative and discriminative. When words are used as indexing features for SCR, they must clearly also be accurate and efficient. We include the list here because it is particularly important to consider each of these points when designing a subword indexing feature inventory for SCR. The lexical word is well entrenched as the default indexing unit — it is not only convenient, but it has also established its usefulness in practice. The lexical word is a "safe" choice since it is always quite reasonable to assume that the semantics of a speech stream is largely conveyed by the speaker's choice of lexical words. Also, lexical words are distributed in length and frequency in such a way that many of them can be modeled relatively readily and thus can be more easily recognized by an ASR system. Moving away from using words as the basic indexing unit, especially in languages like English that are not syllable based, means adopting a model of speech whose architecture is less constrained and less clearly intuitive.

During the design of an SCR system, it is important to keep in mind the basic reliability of word-level features and not to integrate subword information to an extent greater than that which is necessary to confront the OOV problem. Subword units should be used in a way that exploits their ability to represent the speech signal, without sacrificing discrimination, accuracy or efficiency.

There are different methods for using ASR to obtain subword indexing units from the speech signal. The most straightforward manner is to make use of the units that are also used by the ASR system:

- *Acoustic units.* Using the indexing unit as the acoustic modeling unit in the ASR system
- *Lexicon units.* Using the basic vocabulary item in the lexicon (dictionary) of the ASR system as the unit

Phones are examples of acoustic units (in the case of a recognizer that uses the phone as its basic acoustic unit and also a phone-based language model) and words are examples of lexicon units (in the case of LVCSR). Additionally, as mentioned previously it is possible to derive units from post-processing (i.e., re-tokenizing) the output of an ASR system. Here, there are two main approaches:

Units derived from post-processing ASR output

- *Units via agglomeration.* Deriving the unit by merging units in the ASR transcripts during a post-processing step (e.g., extracting phoneme sequences from phoneme-level transcripts)
- *Units via decomposition.* Deriving the units by decomposing units in the ASR transcripts during a post-processing step (e.g., syllabifying word-level transcripts to derive syllable-level indexing features)

In the remainder of this section, we discuss the different levels of subword indexing units that have been used for SCR.

4.4.1 Phone-sequence Indexing Features

At first blush, it seems rather surprising that short phone sequences (phone n-grams) are capable of capturing sufficient meaning in order to support SCR. Generally, semantics in human language is considered to be situated at the word level and higher. However, recall from *Automatic Speech Recognition* that the morpheme is the basic unit of meaning in speech, and that this unit is (often) smaller than a word. Single phones cannot be expected to encode much, if any, semantic information, but phone sequences of length two (phone bigrams) and especially of length three (phone trigrams) and higher grow close to morphemes in length and for this reason it is also plausible that such units can be used to capture semantic information.

Indexing features consisting of sequences of phones were initially proposed by [241] and tested on a small corpus of German-language radio news. The phone transcripts are generated by a phone-based ASR system that uses a phone-bigram language model. The phone sequences extracted from the phone-based ASR transcripts to be used as indexing features are maximally overlapping phone sequences, 3–6 phones in length. The sequences are chosen by a method that eliminates both very frequent and very infrequent sequences. To perform retrieval, the query is first mapped to phone sequences and the VSM is used to compare the query and the documents in the collection. The system described in [254] was an early prototype that made use of this approach, adopting triphone indexing features.

In [304], phone sequences are created by decomposing word-level transcripts generated by a word-based ASR system into phone-based transcripts with the help of a phonetic dictionary. All phone sequences of 3–6 phones in length are used as indexing features. Queries are also converted into phone strings via the dictionary. It is important to point out that the dictionary used for this conversion process is larger than the lexicon of the ASR system. If it were not, the phone-based method would not help to compensate for OOV. Again, the VSM is used for retrieval. The method was tested on a collection of English-language news stories. The use of phone-strings in [304] is intended to emulate the effect of wordspotting in phoneme lattices, which is computationally a more expensive technique. In [174], the method was shown to perform well for English-language broadcast news retrieval, which used phone 5-grams that overlapped by four phones.

In [195], different methods for extracting overlapping phone-sequence indexing features for SCR are explored in detail. This article arrives at the general conclusion that phone-based retrieval is not as effective as word-based retrieval, but there are certain situations where it is appropriate. Specifically, phone-based retrieval is effective for addressing the OOV problem. Further, if speech recognition must be performed on a platform with limited capacity (i.e., a hand-held device), then a small language model, such as a phoneme bigram model, makes the ASR system lightweight and compact. The authors of [195] find that in terms of phone-sequence-based indexing features, a combination of phone 3-grams and 4-grams proved most effective. This result confirms the findings of [304] that phone-based features derived from word-level transcripts are able to help compensate for word-level error. Further, [195] shows that ignoring word boundaries when extracting phone-based features does not affect retrieval performance significantly.

Similar results are achieved by [197], which investigates a wide variety of different subword indexing terms derived from speech transcripts produced by a recognizer with phoneme-level acoustic models and a phoneme-bigram language model. Retrieval experiments were performed using the VSM and 50 topical queries. Overlapping phone trigrams yielded the best retrieval performance. The authors conclude that the overlap of the strings is important because it provides more opportunities for a partial match to be made between the query and the ASR transcript.

Experiments reported in [197] start with phoneme monograms and gradually increase the length of the phoneme sequences. These reveal that retrieval performance first increases and then falls off. This behavior clearly demonstrates the importance of using indexing features that are specific enough to be representative, but not overly specific to the point where they fail to generalize.

We close the discussion on phone sequence indexing units by mentioning work that makes use of sequences of units that are approximately phones. In [87], specialized indexing features are used that are defined as “the maximum sequence of consonants enclosed by two maximum sequences of vowels at both ends.” Note that these sequences do not overlap, but rather the speech signal is cut in the middle of a vowel. A vowel is easily identifiable within the speech signal and also more stable than a phoneme transition, making this choice of segmentation boundary a natural and robust one. The indexing features are extracted from the speech signal using a keyword spotting system. In subsequent work, [241], the authors point out that using phone-sequence features instead of specialized phone-like sequences is beneficial because it greatly increases the ease of feature extraction.

A similar observation has been established in the area of spoken content classification. In [286], two topic classification systems are tested that use phone-sized units as their indexing features. The first uses codebook class sequences (CCS), which are sequences of phone-sized (80 ms) units that have been created by an automatic vector quantization process. The second uses phones generated by a phone-based recognizer. Both feature sets are demonstrated to be suitable for topic classification and in both cases sequences of three units were the top performers.

In [192], experiments are presented that demonstrate the relative benefit of extracting phone-string features from lattices rather than from 1-best phone-level recognizer output. This approach bears affinity to the techniques for detecting spoken terms that will be discussed in subsection 4.6.

4.4.2 Syllable and Morpheme Indexing Features

The classic application of syllable-level indexing features for SCR is in syllable-based languages. Chinese is a typical syllable-based language. An inventory of 1,345 "tonal syllables" covers all the pronunciations for Mandarin Chinese [282]. Chinese is written as a string of characters and, in contrast to Western languages, does not use whitespace to delimit words. Each character is monosyllabic and can correspond to different syllables depending on its word-level context. If a Chinese ASR system uses a syllable-level language model it can easily achieve complete coverage — there is no OOV problem for Chinese syllables. In [15], a collection of 1,000 voice messages are indexed using a syllable spotting approach. Each message is represented by a vector of syllable-bigram features — each component consists of a weight with an acoustic score component and an inverse document frequency (idf) component. Retrieval is carried out with the VSM.

In [282], an ASR system with syllable language models and acoustic models based on half-syllables is used to generate syllable-level ASR transcripts for a collection of Mandarin Chinese broadcast news. Two types of syllable-based features are extracted from the transcripts and are used to index the collection. The first is single syllables and the second is overlapping syllable bigrams. The authors point out the "homophone problem" with Chinese syllables. The same phonetic syllable can correspond to multiple Chinese characters and the characters can play a part in multiple Chinese words. Syllable bigrams help to compensate for this problem by retaining some of the context of the syllable. They consider the overlapping syllable bigrams to be a special case of overlapping phone-sequence features used by [241].

In [41], the discriminative ability of the syllable is explored in detail. Subword-based approaches are shown to be better than word-based approaches for Mandarin SCR. A very practical reason for preferring to use syllables as indexing features is that it is possible to circumvent the need to tokenize Chinese text into words in order to train a word-level language model for the ASR system. However, there are other reasons why subword features perform so well. As pointed out by [41], syllables or characters in Mandarin are capable of covering cases that are difficult for word-level features. The flexibility of subword features aids matching in the case of expressions for the same concept containing slightly different characters. Such cases include transliterations of foreign words, which may choose related but not identical representations for the same string of sounds, and abbreviations, which make use of a subset of characters in the full character sequence used to represent the concept. In the case of perfect transcripts, character monograms perform better than syllable monograms, since they are semantically more specific. However, when error is involved, syllables come into their own. In all cases, subword monograms are to be preferred to word monograms. The best syllable-level indexing features were sequences of 1–3 units in length. The authors of [41] point out that this set of indexing features includes much word-level information, since words are often two syllables in length.

The approach presented in [215] brings together techniques for exploiting multiple hypotheses and subword techniques by proposing subword-based approaches building on word-based approaches that use confusion networks and PSPL lattices. Experiments on Mandarin Chinese show the ability of these approaches to improve over word-based baselines and also examine trade-offs between index size and performance.

The use of syllable-based indexing features for Chinese SCR has been an important source of inspiration for a wide range of feature extraction techniques that eschew the lexical word as the basic feature and opt for smaller subword features. As research into subword-based SCR matured, it became clear that it is not necessary for a language to be syllable based in order to carry a sufficient amount of semantic information on the subword level for the support of SCR systems.

An interesting case is German, which is not considered to be a syllable-based language. The largest possible German syllable is eight phones in length, theoretically admitting the possible existence of millions of syllables. Although the vast majority of possible phone combinations are not used as syllables, the net result is that the syllable inventory for German is for all practical purposes infinite. However, a relatively small syllable inventory still manages to give good coverage. In [155], the same (syllable-level) ASR performance is achieved using a 5,000 syllable trigram language model or a 91K word bigram language model. The syllable recognizer uses phones as the acoustic unit and an inventory of phonetic syllables. Retrieval experiments are performed on syllable transcripts generated by the syllable recognizer and also created by decomposing the word output of the 91K LVCSR system into syllables. In the experiments, syllable bigrams outperform words as indexing features, demonstrating the ability of syllables to provide a flexible fit between the speech signal and the query. Slightly better performance is achieved by using syllable features derived from the word-level transcripts rather than syllables derived from the transcripts generated by the syllable recognizer.

The usefulness of syllable-based features for German-language SCR has a very intuitive explanation. As previously mentioned, German is a compounding language and new compounds can be created, for example the word *Silbenspracherkennung*[2] is composed of *silben* (compounding form of *Silbe*, “syllable”) + *sprach* (compounding form of *Sprache*, “speech”) + *erkennung* (“recognition”) and would not be contained in a standard dictionary of the German language, but is very natural to use in the context of discussing syllable-based methods for ASR

[2] Nouns are capitalized consistent with German spelling convention.

and SCR. The vocabulary of a syllable-based ASR system contains the components necessary to compose this word “on the fly.” Even a LVCSR recognition could come reasonably close by recognizing *Silbe* + *Spracherkennung* or *Silbe* + *Sprache* + *Erkennung*. When decomposed, these sequences would share many syllables in common with queries containing the words *Silben* and *Spracherkennung*.

This explanation for the usefulness of syllables, does not readily extend to English, which is not a compounding language. However, as noted by [197], syllables capture constraints on the combinations of phones that can occur within words (i.e., so-called phonotactic constraints). The syllable level, located between the phone-level and the word-level, provides a balance between capturing incidental patterns (a danger of phone sequence indexing features) and capturing only overly specific features (a danger of word-level indexing features). In [197], syllable indexing features were outperformed by overlapping phone sequence features. However, since longer syllable strings were not taken into consideration, this approach cannot be considered to have exhausted the potential of syllable-level indexing features for English-language SCR.

Researchers have also investigated the usefulness of indexing units that are syllable-like, but not strictly speaking syllables. Such units hold an intuitive appeal, since they strive to combine the strength of phone-sequence indexing features with features that are motivated by frequent occurrence patterns within language. In [197], phone multigrams, defined as non-overlapping, variable-length phone sequences, are compared to other indexing features. Like syllables, they are outperformed by overlapping phone sequence features. Their intuitive appeal does not translate into system performance gains. A similar result is achieved by [174], in which experiments with an inventory of “particles” are conducted. “Particles” are within-word sequences of phones that are automatically derived from a phonetic decomposition of the language model training corpus. The selection of particles involves a greedy process that seeks to maximize the probability of the resulting particle-language model generating withheld text data. The use of particles as indexing features achieves good performance, but is outperformed by the retrieval system that used phone 5-grams as indexing features.

Recent work on syllable-like units is reported in [308]. The authors note that an ASR system might mis-recognize a linguistic unit spoken in the speech signal, but still generate a characteristic pattern in the n-best list. By matching fragments of n-best hypotheses, identical speech segments can be found, without requiring a correct first best ASR system output. Vectors of alternative syllable hypotheses are used to represent each recognized unit, which are then represented in a transformed space where they are quantized to produce a codeword. The approach is evaluated on Mandarin Chinese speech annotations of a photo collection.

Finally, morphemes have also been used as indexing units. Recall that morphemes are the smallest meaning-bearing unit of language. They can be syllable sized or even word sized. Their benefit is the fact that they correspond to the modular units that are used to morphologically derive the language. These units are used for morphologically complex (highly inflective) and, in particular, agglutinating languages such as Turkish and Finnish. An ASR vocabulary of a size that would give good coverage in a language such as English, yields high OOV rates for such languages, since it cannot cover all the morphological variants of the individual base words. In [271], a set of morpheme-like units, derived via a statistical process, is used for SCR of Finnish-language read news stories. The morpheme-level subword units are produced directly by the ASR system and serve to limit the size of its lexicon and language model. The morpheme inventory can be created by using linguistic decomposition rules, or it can be created with a statistical approach, as in [271]. In [8], morpheme-based units are used for both ASR and STD on a corpus of Turkish broadcast news retrieval.

In [271], the morpheme-based output of a Finnish language speech recognizer is transformed into WCN form using the algorithm. The VSM with *tf–idf* weights is used for retrieval. The *tf* factor is substituted with a lattice specific score, which is calculated in one of two different methods. The first method sums the posterior probability of the morpheme over every occurrence in the confusion network. The second method ranks the morphemes at each time point by their posterior probability and calculates the *tf* factor by summing the inverse of the rank position over every occurrence of the morpheme within the confusion network. Both methods are reported by the authors to improve retrieval over the use of 1-best transcripts. A priori, the second method is better motivated than the first. The posterior probability of a unit can be compared with other units recognized at the same time point. For these reasons, a ranking of alternatives at a certain time point in a lattice can be anticipated to be relatively reliable. It is more difficult to compare units that the recognizer has hypothesized at different points in the signal. Such comparison requires appropriate normalization of the posterior probability, an aspect of the approach not detailed in [271]. The experimental results reveal that tighter pruning of the confusion network leads to better performance. Pruning, in this case, has the effect of introducing a threshold on the recognized terms that are admitted for indexing. The *tf–idf* model used here is a simple linear *tf* function. Without comparison with a more sophisticated term weighting model, and also comparison with a simple thresholding of term hypotheses based on confidence scores, it is difficult to conclude the general contribution of posterior-probability-based weighting to the improvement of confusion network-based SCR.

4.4.3 Moving Beyond Exact Matches

Thus far, we have seen that subword indexing units have the ability to compensate for ASR error. The modularity of subword units allows units from a relatively small inventory to be assembled on the fly during the speech recognition process to fit to the speech signal. In this way, subwords have the potential to cover OOV words and also to compensate for unexpected acoustics, such as pronunciation variants. Techniques discussed up until this point extract features from ASR transcripts by using exact matches. In other words, if the feature does not match the transcript perfectly at a given point, it cannot be extracted from the transcript at this point. Some approaches have introduced additional flexibility into subword indexing techniques by loosening this restriction.

There are two basic ways in which an approximate match between the indexing term and the ASR output can be used for indexing term extraction. The first method is to go beyond the 1-best list of the recognizer and use a lattice that encodes alternative recognition hypotheses. Indexing methods that pursue this approach will be discussed in subsection 4.6, together with other lattice-based approaches for detecting spoken terms. The second approach makes use of the 1-best transcripts of the recognizer. Instead of explicit knowledge about possible recognizer mistakes encoded in a lattice, these methods impose a simple model of speech recognition error and use that model to calculate fuzzy matches between indexing terms and speech recognition transcripts. If a feature is very similar to the 1-best ASR transcript at a given point, then it is extracted at this point. In general, similarity is determined by the minimum edit distance (MED) between the transcript and the feature at the point in question. The MED is the minimum number of insertions, deletions or substitutions that are required to transform one string into another, and provides a useful reflection of similarity. It is calculated using dynamic programming techniques. In the case of feature extraction for speech indexing, the strings that are compared are usually phone-level strings, that is, both the transcript and the indexing feature to be extracted are represented in terms of their constituent phones.

Such "fuzzy matching" techniques for feature extraction should be deployed with great care in an SCR system. Their advantage is the ability to cover more of the speech signal by introducing an improved robustness to ASR error. Their disadvantage is that they admit too many matches with the speech signal, that is, they often identify similarity where no underlying similarity exists. A difference of only one phone can correspond to a large semantic difference, cf. "kill" vs. "fill." Even if the range of possible differences is constrained to phones that are acoustically similar and thus prone to confusion by the recognizer, similar phone strings correspond to wide semantic differences: "die" vs. "tie." Whether or not inexact matches should be used will depend on characteristics of the domain of application and the use scenario. For example, if different information needs of users correspond to spoken content with clearly different vocabulary use, then inexact matches may not be as damaging. Further, if the spoken content to be retrieved takes the form of longer documents, the large number of indexing features extracted from these documents may compensate for a few unreliable cases.

A technique for fuzzy matching on phoneme transcripts was introduced by [288, 289], as mentioned above in the context of confidence scores. Here, a slot detection approach was applied to identify putative query word occurrences in phone transcripts. Slot detection compensates for ASR errors by admitting exact matches.

In [201], similar phoneme sequences were mined from phoneme transcriptions and used to categorize a set of unseen speech documents. The inexact matching technique worked well for this rather limited domain. This result contrasts with SCR work such as [195], where expanding the query with typical errors and close phone strings was shown to degrade the performance of phone-sequence-based retrieval.

Further experiments in [195] pursued a more constrained approach. Phoneme classes were built by grouping the phonemes in the recognizer's acoustic model inventory into 20 broad classes according to acoustic similarity. These classes capture phonemes that are potentially easily confusable by the recognizer, introducing an exact match element designed to compensate for recognizer error. This approach was shown to be robust and its performance was closely followed by that of 4-grams built of phoneme classes.

We close our discussion of indexing units that make use of inexact matches by mentioning that many of the same effects can be built into the SCR system at query time. In [155], syllable-level fuzzy matching techniques are shown to improve over exact match techniques for German-language spoken document retrieval. This approach differs from other inexact match techniques in that the match is calculated at two levels: first, the phone-strings of syllables are compared to calculate a syllable score and then syllable strings are compared to determine the presence of an indexing feature. The approach is designed to compensate for speech recognition errors that involve the compression of longer words into shorter ones — an effect that arises due to the fact that the standard HMM-model ASR framework does not effectively model the length of words. Note that this approach effectively exploits a simple model of ASR error expected in the ASR transcripts.

Another query-time matching technique involves query expansion. In [173, 174], a method is proposed that expands the query with a list of confusable phrases. In effect, this approach attempts to make a query-time “guess” at the way in which the ASR system has mis-recognized terms in the query.

In [257], a probabilistic form of term weighting is used that makes use of information on phone confusion. The method works well, but the authors note that for the particular application (a distributed learning application) the small size and constrained topic could be what makes the phoned-based approach effective. In the next subsection, we turn to techniques that propose combinations of approaches for improving SCR performance.

4.5 Hybrid Approaches for SCR

As discussed at the end of *Automatic Speech Recognition*, the higher goal is to strike a balance between ASR and indexing. The ideal system should retain the advantages (i.e., modularity, flexibility) associated with isolating ASR and search components, while striving to retain information important for indexing. The tightest integration between the ASR system and indexing is achieved when the identities of the indexing features are known ahead of time and can be used as units within the ASR system. In most systems, they would be used as units in the language model, but it is also possible to use them as acoustic units, if they are known in advance. It would make sense, for example, to consider using syllables as acoustic units in a syllable-based system. In practice, it is not possible to have complete knowledge of necessary indexing features at ASR time, in particular due to the OOV problem. A very loose integration of ASR and indexing involves treating the recognition output as a given and applying a post-processing step in order to extract useful units from this output. A middle road is to use the ASR system to generate an intermediate form of output that is optimized to be used later for the extraction of features on the fly at search time.

Since the beginning of research and development in SCR, the lexical word has been widely used as an indexing feature and has repeatedly proven its value. The word represents a natural choice, not only because it is the canonical meaning-bearing unit in language, but also because it is convenient and text IR also offers many models that have been proven to work well using word indexing features.

The early speech retrieval systems, in particular [131, 181, 231], used word-level features derived by a word spotter in order to perform classification and retrieval of spoken documents. A word spotter takes as input a list of words and the speech signal and outputs points in time at which words appear. The obvious limitation of this approach is that the indexing vocabulary must be known before the spoken content is processed by the ASR system. We return to further discussion of word spotting, as well as vocabulary-independent methods of extracting word-level indexing features from ASR output, in subsection 4.6.

With advances in Large Vocabulary Continuous Speech Recognition (LVCSR), word-level transcripts became a valuable source of word-level indexing features for SCR [121, 132, 303]. The advantages of using word-level transcripts are twofold. First, an LVCSR system uses a word-level language model spanning a context much larger than that covered by phoneme-based recognizers. Second, word-level transcripts can be indexed in a relatively straightforward manner by simply applying conventional IR techniques. The disadvantage of using LVCSR transcripts is the OOV problem. A word-level transcript can never contain a word form that is not explicitly included in the vocabulary of the ASR system.

We have seen that subword units can be used as an indexing unit in an SCR system in order to address *The challenge of covering all possible words*. However, they must be deployed with care. The downside of their robust and flexible matching capability is their ability to introduce into the system spurious matches between spoken content and query terms, leading to a drop off in SCR precision. A natural means of integrating subword indexing units into an SCR system is to take a hybrid approach that makes use of indexing features of different levels. In such a way, the system is able to exploit the flexibility of subwords (such as syllables), but also the reliability of larger units (such as words or phrases). There are three different ways in which indexing units of various levels can be integrated into an SCR system: *Low-level integration*, *High-level* *integration* and *Query-selective integration*. In the following discussion, we treat each in turn.

Low-level integration. Low-level integration approaches involve extracting indexing features of multiple types and lengths and combining them into a single index. Intuitively, the issue that arises is potential duplication of the same information within the index. However, above we saw the advantage of information duplication in the case of phone-sequence features: overlap of features helped to improve performance. Because an SCR system must deal with errors in the transcripts, it is not possible to know a priori if duplication of features is introducing redundancy or helpful, new information.

Since the early days of SCR research, low-level integration has been used to combine different levels of indexing features. In [120], a combination of features derived from two speaker-dependent ASR systems was used for the task of retrieving read English-language news reports. One system generated word-level transcripts and the other phone lattices. Query words not present in the word-level transcripts were scanned for in the phone lattices at search time, using the *wordspotting in phoneme lattices* technique. The combination led to a sizable improvement in retrieval performance.

In [132], speaker-independent versions of the same ASR systems were used for the task of retrieving English-language voice messages. The combination yielded a small improvement over either feature source used in isolation. The experimental results were slightly better for the system that made use of both word-transcript and phone-lattice derived indexing terms, independent of whether the query word was present in the vocabulary of the ASR system that produced the word-level transcripts. This result suggests that it is advisable to risk information duplication when selecting indexing features, rather than risk leaving out information.

Since this initial work, other approaches have also made use of low-level feature integration. In [195], an improvement is achieved on an English-language broadcast news SCR task by combining phone sequence 3-grams and 4-grams. Similarly, in [197], a small improvement is reported for English-language broadcast news retrieval when phone-sequence features of various lengths were combined. The combination improves only marginally over the best scoring individual phone-sequence feature (phone trigrams). The authors comment that the small improvement may be due to the fact that the same relevant documents are scoring the same for the different representations. Comparable results have been obtained in the area of spoken document classification. In [60], German-language spoken document classification is performed using a Support Vector Machine and a combination of different sorts of features, and a small improvement is achieved with the combination of word and character trigram features.

High-level integration. Integration at a high level involves querying multiple indexes, one for each sort of indexing feature, and combining the different results lists returned. A common method uses a linear combination of the retrieval scores of results in each list. High-level integration became a firm tradition in the early days of SCR. In [132], document retrieval scores returned from searches using word-level ASR transcripts and from searches using phone lattice spotting *wordspotting in phoneme lattices* were linearly combined and the combined scores used to generate the final ranked list of spoken documents. The high-level integration approach yielded a small improvement over either individual approach.

In [304], a mixture of word-level indexing and phone-sequence indexing is used for English-language broadcast news retrieval. The use of phone-sequence features alone did not yield the same performance as word-level features. However, a linear combination of word-based and phone-sequence based retrieval outperformed either individual system. Subsequent work on English-language broadcast news SCR has also taken a high-level integration approach for combining features. In [197], a linear combination of results generated using different length phone sequence indexing features yields an improvement. The weights with which the different retrieval scores entered the linear combination need to be correctly set to favor better performing units. If they are not, no improvement is achieved. In [172, 174], a substantial improvement is achieved by a system that uses a linear combination of word-level retrieval scores and retrieval scores from an index using phone 5-grams. Again, it is important to optimize the weights used in the linear combination.

In [41], Mandarin Chinese SCR is shown to benefit from the addition of character or word information to syllable level results. Retrieval results from word and subword scales are merged via linear combination to improve cross-language English-Mandarin retrieval in [170]. Linear combination of retrieval scores has also been shown to yield improvement in [167], where an optimal weighted combination of scores generated by subword units of various sizes yields clear performance improvement.

In [209], a linear combination of scores from a ranked utterance retrieval system and a system that uses word-level ASR transcripts is proposed in order to provide a robust approach to the OOV problem. A training set is used to learn effective mappings between the scores and probabilities of relevance to the user query.

Linear score combination has been shown to yield good performance when the results retrieved by the contributing retrieval methods (in this case, retrieval with respect to different sets of indexing features) are of high quality [276]. Although linear combination of retrieval scores is widely used in SCR, we would like to point out that there are many other possible methods for score combination that have been developed in the text-based IR literature (cf. e.g., [19]). Due to the large amount of ASR transcript noise faced by most SCR systems, other methods of high-level integration most likely have good potential to further improve retrieval performance.

Query-selective integration. This method analyzes the query and chooses the index to be used for retrieval on the basis of this analysis. Most notably, this method is used in order to deal with OOV query words. An OOV query word is a word occurring in the query, but which cannot possibly be found in the word-level ASR transcripts because it is not contained in the vocabulary of the ASR system. In practical applications, the vocabulary of the ASR system is known to the retrieval system, or can be surmised by simply compiling a list of the word forms that occur at least once in the available speech recognition transcripts.

In [172, 174], it is noted that using phoneme sequences rather than words as indexing features increases recall, but also increases the number of false alarms. A better balance is achieved by a system that queries the word-level index if the query word is in the recognizer vocabulary and the phoneme-sequence index otherwise. The resulting performance is on par with high-level integration involving a linear combination of scores, but exhibits two advantages. First, the number of false alarms remains low and second, it is not necessary to optimize the weights in the linear combination. Although this experiment suggests that query-selective integration is a promising approach, it does not yet completely solve the problem. The queries used in [172, 174] are overwhelmingly only one word in length, meaning that the system avoids difficult decisions about which index to use. This experiment bears strong resemblance to an STD task. Although the VSM is used as retrieval, a document is judged to be relevant if it contains the query term.

Further, query expansion and document expansion are techniques that can be used to improve the match between query and spoken content or the suitability of weighting scores. Assuming that the use scenario of the system is compatible with expansion, any attempts to exploit confidence scores or word lattices should aim at going above and beyond the improvement that can be achieved using expansion techniques. These techniques are discussed further in *Spoken Content Retrieval beyond ASR Transcripts.*

4.6 Techniques for Detecting Spoken Terms

The challenge of covering all possible words is a tough one. In the mid-1990s, a basic strategy emerged in the work of [120, 132] that remains today a framework for an effective way forward in addressing the OOV problem. This strategy involved splitting the process of generating indexing terms from spoken content into two steps, the first occurring at indexing-time and the second at query-time. For SCR, indexing involves the use of an ASR system to generate speech recognition transcripts, computationally a very intensive process. If the ASR system can be used to generate an intermediate representation, then that representation can be used at query-time to identify the positions of helpful indexing features contained in the query. The two-step approach faces several challenges: first, determining the appropriate intermediate representation to be generated by the ASR system; second, ensuring that the intermediate representation can be stored in a way that makes it efficient to search at query time; and, third, ensuring that the process of identifying indexing terms at query-time does not give rise to spurious terms that negatively impact SCR performance. The two-step paradigm experienced a resurgence of popularity in 2006 with the introduction of the concept of STD by the NIST STD task [75]. Recall from *Overview of Spoken Content Indexing and Retrieval* that there is a close relationship between STD and the core SCR task, relevance ranking. The core SCR task involves using ASR transcripts to find speech media that is topically related to the information need underlying the initial query. In use scenarios in which the information need of the user is to find mentions of particular words or phrases within the speech media, the search problem reduces to STD and the IR model used is the most basic model conceivable, namely a Boolean search. In this section, however, our main interest in keyword spotting and STD is their usefulness in extracting indexing features that can be used to tackle the core SCR task. Note that many of the techniques that have been discussed above for SCR (e.g., use of lattices and subword units) also play an integral role in keyword spotting and STD.

4.6.1 Wordspotting Using Fixed Vocabulary

In the original work in the area, wordspotting is defined as finding occurrences of a specific list of words or phrases within a speech document or a speech stream and we adopt this definition here. In contrast to later versions of this task (e.g., keyword spotting and STD), in the wordspotting task it is assumed that the identity of the words to be located is known to the system already at ASR-time and used by the system as it processes the speech signal. The dawn of the wordspotting era was arguably 1990, with the publication of [232], which described an HMM-based wordspotting system. These approaches are aptly referred to as “voice grep” by [241]. Such systems were proposed also in order to handle voice input [301]. As previously mentioned, the application of wordspotting systems to tasks related to searching speech focused on message classification, e.g., [230]. As commented by [264], they remain appropriate for monitoring tasks, such as broadcast news monitoring.

4.6.2 Keyword Spotting in Phoneme Lattices

Wordspotting is limited in its practical application since the words to be extracted from the speech signal must be known to the system at ASR-time. The keyword spotting paradigm removes that restriction, thereby directly addressing the OOV problem. The first approach to vocabulary-independent query-time word location was proposed by [122]. The approach involves generating a phone-lattice at ASR-time and searching this lattice for instances of query terms at query time. This approach was applied to SCR by [27, 28, 120, 121].

When a linear scan of phone lattices is used, wordspotting is practical for small collections, but does not scale effectively when the quantities of spoken content to be indexed grow larger. The task of determining matches between this ASR output and the query words is then taken over by other components of the SCR system. In the wake of the initial work on keyword spotting in phoneme lattices, two lines of investigation were developed that focused on exploiting this basic principle.

In [72], an approach is proposed that makes use of synchronized lattices. All alternative hypotheses are synchronized with the first best phoneme hypothesis together with their posterior probabilities. The synchronized lattice is relatively computationally cheap to sweep and also outperforms the first best phoneme hypothesis.

Another opportunity for exploiting lattices, which deserves mention, is lattice matching. In [283], a method for matching a lattice representing the query with lattices representing the spoken documents in the collection is proposed.

4.6.3 Spoken Term Detection

STD addresses basically the same task as keyword spotting, namely to find spoken instances of terms in the speech signal that were unknown to the system at ASR-time. Research on STD was launched by NIST in

2006 with the Spoken Term Detection Evaluation Plan [199]. There is not a sharp boundary between keyword spotting and STD, but in general STD approaches are distinguished by their emphasis on speed and scalability, that is, approximating the sequential lattice scan typically used by keyword spotting approaches. This distinction is reflected in the four requirements for STD identified in [243]: high recall, high precision, time and space efficiency (i.e., scalability), and flexibility (i.e., adaptability of the system to new conditions). These requirements are based on an examination of real-world use scenarios. The relative importance of each varies from scenario to scenario. Effective approaches to STD will be able to achieve the appropriate balance in these requirements.

In general, STD approaches involve an initial decoding stage that generates an intermediate representation of the search space. This representation seeks to extract the information from the speech signal necessary to be able to identify the occurrence of terms yet unknown at the time of the initial decoding. More detailed information on STD technology is available in [75]. In this subsection, we briefly cover some more recent highlights of the field.

An approach called "Dynamic Match Lattice Spotting" (DMLS) is presented in [264]. This method uses an inexact, MED match to search the lattice in order to compensate for phoneme recognition errors. In this way, it aims to compensate for phone strings not included in the phone lattice, which the authors cite as a major shortcoming of lattice-based approaches. Since the dynamic match is an inexact match, it can be expected to generate false alarms. False alarms are controlled by setting a threshold on the match score and also by carrying out a second, verification step. Speed-up of lattice search is achieved by limiting the length of the search string (i.e., query phone-string) and storing for every timepoint (i.e., node) in the lattice all phone-strings that begin at that node.

In [263], a comparison is made between the use of acoustic units based on graphemes and units based on phonemes for STD. The use of graphemes allows the system to circumvent the step of converting the queries from a string of characters to a string of phonemes, which can be a particular weak point in STD in the case of OOV queries (queries falling outside the available pronunciation dictionary).

4.6.4 Spoken Utterance Retrieval

The task of spoken utterance retrieval (SUR) also involves matching occurrences of query words with their mentions in the speech signal. Here, the collection is segmented into short audio documents referred to as spoken utterances. In order to be relevant to a query, a document must contain a mention of the query words.

In [252], a two-step approach is proposed in which promising lattices of a select group of speech segments (each up to 30 seconds in length) is first identified using a rough match and then scanned using an exact match.

The relative performance of lattice and confusion network approaches applied to SUR is evaluated and discussed in [141]. The authors also use a set-of-words approach, which destroys the timing and overlap relationships between the links in the lattice, treating it as a bag-of-words. This method is computationally very cheap and surprisingly effective, especially in high WER settings. This result suggests that it is important to check the performance of simple, computationally light techniques before unleashing sophisticated lattice calculations to compute effective term frequencies.

In [210], SUR is approached as a "Ranked Utterance Retrieval" task in which speech segments are ranked according to a confidence score reflecting whether or not they contain the query. A fuzzy matching approach is implemented by degrading queries using phone-based confusion, phrase-based confusion statistics and confusion based on factors consisting of phone classes, with best performance yielded by the latter.

4.6.5 Hybrid Approaches

The most effective approaches to SUR and STD use multiple hypotheses of the ASR systems (i.e., lattices) and also exploit a hybrid combination of word-level and subword-level units. A hybrid approach to SUR was proposed by [238], who characterize it as a generalization of the combination of word indexing and keyword spotting in lattices, originally used by [132]. Within the lattices, the score of a lattice link (either a word or a phone) is calculated by counting the number of times the word occurs on a path and multiplying by the probability of the path. As the authors observe, this score is a lattice-based confidence measure. Three different strategies of combining the word index and the phone index were investigated: combining the results returned by both, searching the phone index in case of OOV, but otherwise using the word index, and using the word index if no result is found in the phone index. The latter strategy returned the best results. Use of word lattices achieved a 5% relative increase in performance over words alone. Use of word and phone lattices achieved an 8%–12% increase. The experiments demonstrate that use of lattices is more beneficial in the case of spontaneous speech (here, teleconferences) than it is in the case of planned speech (here, broadcast news). In [114], an approach to SUR is presented that makes use of a hybrid confusion network containing both words and phones.

In [312], a related approach is proposed and applied to both SUR and STD. This approach makes use of a lattice indexing technique aimed at improving the speed of lattice search and thus the scaleability of the algorithm. Spoken utterances are retrieved using an Expected Term Frequency (ETF), which estimates the effective number of occurrences of the query term that they contain by multiplying the expected number of words in the lattice by the probability of the query occurring in the lattice. The probability of a query being observed in a lattice is estimated by taking a product over the n-gram probabilities that have been estimated on the lattice for each phone in the query. The index then only needs to store counts of phone strings occurring in the lattice. ETFs for words and phones are integrated using a linear combination. STD is accomplished by carrying out a detailed match on segments with high ETF values in order to localize the position of the query within the segment.

In addition to phones, syllables have also been combined with words to create hybrid STD systems. In [184], two hybrid approaches for STD that combine syllables and word ASR output are proposed. The first approach combines words with fuzzy search in a 1-best syllable transcript. This approach is outperformed by a second approach that combines words with an exact search on a syllable lattice.

In later work [185], the authors investigate a different variety of hybrid STD system that merges the two syllable search spaces: the fuzzy search on 1-best syllables and the exact search in the syllable lattices. Although some instances of spoken terms are found by both approaches, there are also cases that are found by one technique and not the other. The combination of the two subword-based techniques was shown to achieve additional gains. The authors note that the lattice matching helps to compensate for ASR error and the fuzzy search on the 1-best transcript contributes to fitting the canonical syllable decomposition of the query with the syllable pronunciations used by the speakers.

Applying similar reasoning to an SUR-type task, [182] observes that different versions of lattice-indexes and confusion networks make different contributions and proposes a learning-to-rank approach from IR to optimally integrate the two representations. In general, hybrid techniques for STD and SUR and those for SCR (cf. subsection 4.5) can be considered to represent a continuous space of approaches that can contribute to improving SCR performance.

In summary, this section has presented approaches that make maximum use of the output generated by an ASR system. Approaches using multiple hypotheses, approaches using subword units and approaches exploiting combinations have been presented. In the next section, we step back from the recognizer, taking a broader view and looking at techniques that can be used to augment ASR for the purposes of improving SCR.

5

Spoken Content Retrieval beyond ASR Transcripts

The previous sections have focused on the contribution made by automatic speech recognition (ASR) to spoken content retrieval (SCR). Here, we discuss approaches that go beyond the basic indexing of ASR transcripts and make use of other techniques and other information sources in order to improve SCR. Subsection 5.1 is motivated by the need to make use of additional dimensions of speech media. Then we turn to treat the final two challenges from the list of SCR challenges in subsection 2.4. *The challenge of context* is addressed by a subsection presenting techniques for augmenting ASR transcripts and, finally, *the challenge of structuring spoken content* is addressed by a subsection discussing techniques for organizing and representing spoken content.

5.1 Motivation for Moving beyond ASR

The motivation for moving beyond basic ASR is to improve the integration of ASR and information retrieval (IR) within the SCR system. Improvements are achieved both by exploiting techniques for IR that have been demonstrated to be effective in the text domain and also techniques specific to SCR. The driving aim is to take advantage of opportunities to target the areas in which SCR is more challenging than IR or otherwise fails to reduce to IR. In particular, we return to the differences between SCR and IR and the considerations that should be made when combining ASR and IR, first discussed in subsection 3.4. This subsection covers factors that motivate SCR beyond basic ASR.

5.1.1 The Content of Speech Media

ASR transcripts provide a less than fully complete representation of the spoken content of speech media. Their most obvious shortcoming is noise, in terms of recognition errors, that is, the insertion, deletion or substitution of words spoken in the speech stream. However, depending on the domain, the speech stream itself can actually be considered underspecified in terms of the information content it contains. In particular, when SCR moves away from domains involving planned speech, such as broadcast news, speech is frequently informal or spontaneously produced. In such cases, the meaning conveyed by the spoken content has a strong dependence on contextual factors, which are not directly represented in the transcripts, or which, if present, may be more challenging to extract. Such factors include facts about the speakers' immediate surroundings, recent events or specific background knowledge that is shared between the speakers of the specific topic under discussion.

Figures 5.1 and 5.2 give examples of two spoken passages on the same subject, which illustrate the difference. Characteristic differences between planned speech and informal speech can be observed by comparing these examples. In planned speech (Figure 5.1), the content words related to the topic are more frequent and repeated (i.e., "speculaas" and "shortcrust biscuit"). In the informal speech (Figure 5.2), words related to the topic are descriptive rather than exact (i.e., "spice cookie" is used instead of "speculaas"; "windmill cookie" is an inexact reference since not all speculaas are shaped like windmills).

In the informal speech, the context serves to establish and support the topic. Pronouns (they, them) are then used by the speaker to make the connection to the entities under discussion. Other pronominals

Speculaas is a shortcrust biscuit containing spices including cinnamon, nutmeg, cloves, ginger, cardamom and white pepper. In the Netherlands, speculaas is traditionally associated with the holiday of Saint Nicholas, celebrated at the beginning of December.

Fig. 5.1 Example passage of planned speech.

You can buy these spice cookies here starting at around this time of year. They are quite different from the Dutch windmill cookies that we have where I'm from. A lot of the people I know who come to live here don't uh really like them. They're just too, well . . . I guess they are spicy in a strange spicy way. You just have to grow up here eating them.

Fig. 5.2 Example passage of spontaneous speech.

establish the connection to the time ("this time of year") and the place ("here") — the month (December) and the country (Netherlands) are not mentioned directly. Further, the people mentioned ("a lot of people I know") can only be identified by knowing more about the person who is producing the informal speech. An IR system, which relies on word overlap between query terms and the indexing terms representing items, will clearly have less difficulty generating a match score for the planned speech example than the informal speech example. Issues of the contrast between planned and informal speech are explored further in [135].

Clearly, the planned versus informal distinction is not limited to the domain of speech and text IR must also deal with the challenges of informal text, especially in conversational social media such as blogs and microblogs. However, SCR faces the additional difficulty that increased levels of spontaneity lead to an increased level of error arising from the speech recognition process. For example, articulation may be less distinct and, as seen in Figure 5.2, the ASR system faces disfluencies and restarts. In short, the production characteristics of informal speech make it quite challenging to recognize.

In subsection 3.4.3, we discussed how IR models are able to naturally compensate for ASR error, to a certain extent. In the following material, we turn to the discussion of additional techniques that can be deployed to address both ASR error and also the underspecification of informal content.

5.1.2 The Structure of Speech Media

Some speech media are produced as a series of semantically discrete segments, for instance, as a series of reports in a news program. In principle, each of these segments can be taken as a document in the SCR system. However, SCR systems are rarely faced with the presence of a single, readily evident manner of decomposing speech media into units. Spoken content produced in informal settings is inherently less structured than planned content. Moreover, in many cases, there is a mismatch between the topic of spoken content and the units into which it most readily decomposes. For example, a meeting could be relatively easy to segment into speaker turns. However, the topic under discussion may last through multiple speaker turns. In contrast, a podcast may decompose most readily into episodes corresponding to individual audio files, but one podcast can contain multiple topics.

These mismatches create a tension that makes it difficult to determine a single set of ideal units for an SCR system. The system should display result units that are intuitive units for the user, but it should calculate a match with the user query that is maximally topically homogenous, with respect to retrieval units. Neither may correspond exactly to the most easily identified units within the speech media. Another option is to allow the calculation of units to take place dynamically and dependent on the user query. In short, an SCR system stands to benefit greatly from structuring techniques that allow it to deal with the lack of structure, or the lack of optimal structure, that is clearly demarcated in the speech stream. Techniques that move beyond conventional ASR transcripts can provide important support in this area. These techniques will be discussed in subsection 5.3.1. In *Accessing Information in Spoken Content*, we will return to the issue of speech media units for SCR and in particular consider how speech media results can be displayed to the user in the interface.

5.1.3 User Queries for SCR

The basic aim of an IR system is to provide users with items that satisfy their information needs. The *Introduction* stressed that SCR is a "finding content" task, meaning that the system goes beyond locating mentions of the query terms within the speech stream to returning results whose content as a whole is relevant to the information need. A detailed examination of the concept of relevance in IR is beyond the scope of this survey. It is sufficient to consider relevance to be a match in topic or other characteristics between the query and the results that lead to satisfaction of the user information need. The discussion of user information needs for SCR in subsection 1.2.2 mentioned the broad diversity of goals that motivates users to turn to SCR systems. This diversity of user needs in SCR is reflected in the diversity of user queries that faces an SCR system. In this subsection, we show how user queries for SCR provide motivation for going beyond ASR. Figure 5.3 gives a selection of example queries that have been used by retrieval benchmarks.

These queries specify the topic of the desired speech media results. Topics can be seen to range from very broad to quite specific. In addition to specifying what the desired content should be about, users might also include in their search requests other facets of the desired relevant results. For example, they might add an indication of the identity of the speaker, the language or the name of a particular program to the queries. The two examples from TRECVid show that speech can also be interesting for users in a setting that is otherwise focused on retrieving video based on its visual content.

It is good to keep in mind that in SCR, as in text IR, the user query is always an imperfect realization of the user information need. When formulating a query, the user makes a choice of query terms. This choice might be less than optimal for a variety of reasons, including the user not knowing or not recalling the most specific terms, the user having the "bad luck" of having chosen synonyms or paraphrases, or the user leaving the query incompletely specified in an attempt to avoid unnecessary effort. Although explicit models of user query formulation behavior have potential to lead to better matches between queries and results, few studies examine the mapping between queries

`Find reports of fatal air crashes.` (Spoken Document Retrieval Task, Topic 62, TREC-7 [82])

`What economic developments have occurred in Hong Kong since its incorporation in the Chinese People's Republic?` (Spoken Document Retrieval Task, Topic 63, TREC-7 [82])

`Jewish resistance in Europe` (Cross-Language Speech Retrieval, Topic 1148, CLEF [295])

`Find the video just of a man wearing a suite, tie, and glasses, speaking French.` (Known Item Search task, Example Query 12, TRECVid [255] 2010)

`Find the video where a man talks about unions, I think his name was Miller.` (Known Item Search Task, Example Query 19, TRECVid [255] 2010 KIS)

`Masavo X defines voice acting.` (Rich Speech Retrieval Task, Development Set Topic 7, MediaEval 2011 [156])

`Could you find the portion of the talk where they are discussing optimal times for talks?` (Rich Speech Retrieval Task, Development Set Topic 12, MediaEval 2011 [156])

`CNN report on Hillary Clinton.` (Rich Speech Retrieval Task, Development Set Topic 21, MediaEval 2011 [156])

Fig. 5.3 Example queries from SCR benchmarks.

and information needs for SCR. An exception is [21], which examines podcast retrieval. As previously mentioned, this work found that queries containing a person's name ambiguously reflect a user need either for information about the person or for information spoken by that person. Further, the impact of the imperfect recall of users' memories on the queries that they formulate is explored. In particular, the work investigates the difference between cases in which the user recalls the exact quote and cases in which the quote is not recalled exactly and users use only a few key words and possibly add some indication of the speaker (e.g., the difference between “The internet is a pressure cooker” and “Maris internet pressure cooker”).

Researchers have long been aware of the importance of dimensions beyond the spoken content, for example speaker- and language-based queries [202]. The larger vision, formulated early-on in the history of SCR, for example in [275], is that combining different forms of analysis will lead to better SCR systems. However, it is often the case that systems are designed without carrying out user studies to collect information on user needs or on how users formulate queries. The diversity of examples in Figure 5.3 serves to illustrate that it is important to understand what characteristics of spoken content are important for users and how they formulate queries in order to design the most effective possible SCR system. This diversity motivates moving beyond ASR for the design of SCR systems, the topic to which the remainder of this section is devoted.

5.2 Augmenting ASR and Spoken Content

The shortcomings of ASR transcripts can often be addressed by using techniques aimed at extending them to make them more complete or to compensate for error. This subsection presents methods for exploiting metadata that may accompany speech media and then looks at additional expansion techniques based on text IR approaches.

5.2.1 Exploiting Metadata

Metadata are commonly defined as “information about information” and a wide variety of metadata can be associated with speech media. Specific metadata that may accompany spoken content in a particular collection include title, creator, source, names of speakers, date of recording or broadcast, language spoken, descriptive keywords or a content precis or summary and closed captions. With the rise of social media, social metadata such as tags and user comments have become increasingly important. The quality and completeness of metadata depends on their source. The variation is enormous, ranging from very complete metadata records in professional archives to recordings of informal discussion that may have no metadata other than the time stamp reflecting when they were recorded. Whatever metadata are present, however, should be conscientiously exploited by an SCR system. In this subsection, we examine how metadata can be used to support retrieval.

5.2.2 Exploiting Manually Generated Metadata

The value of metadata in SCR depends on a number of factors, including its content and quality as well as the SCR task to be addressed. We discuss use of human-generated metadata by covering a series of example domains in which it has been applied and example systems that have applied it.

The value of metadata, even very simple metadata, was already recognized in the early SCR work. The early Video Mail Retrieval using Voice (VMR) project [27] worked with video messages, which replaced the message of conventional email with a video. Sender and query fields were used to sort retrieved items. Such metadata are simple, but very valuable in narrowing the amount of content that is presented to the user in the results list. Although such narrowing is also convenient for focusing search in conventional email, in the case of video messages, it also saves the user a great deal of time in reviewing results, which must be individually watched.

In some cases, metadata are available because an audio collection has been annotated by hand for a particular purpose. A well-studied example of this case is the collection of oral history interviews from the Shoah Foundation Institute [33, 204]. These data were used in the CLEF Cross-Language Speech Retrieval track in 2005, 2006 and 2007 [205, 217, 295]. The manual metadata here were assigned by domain experts who drew keywords from domain specific ontologies and also wrote short summaries using their knowledge of the domain. Results from the CLEF workshops showed that retrieval effectiveness using only the ASR fields is poor, while incorporating the metadata gives much better performance. The usefulness of the metadata can be related to the challenging nature of the task, which involved spontaneous speech. Many of the query words were not present in the spoken content. Further, recognition was challenging, with recognizer word error rates of around 20%. These experiments made clear the utility of combining indexing fields containing ASR transcripts with indexing fields containing human generated metadata. However, it is important to consider how these fields should best be combined and to avoid the perils of naive approaches. The formal analysis of field combination in [229] provides further details on relevant issues.

Lectures are a domain in which accompanying metadata are often readily available. In [248], a case was investigated in which very few metadata were available, only titles, abstracts, and bibliography of the speakers. Metadata still provided improvement over using ASR transcripts alone. Performance with metadata alone was not satisfactory. Adding speech content improved the retrieval performance by 302% (relative).

A more detailed source of information associated with lectures is slides. When slides are available, even highly error-filled lecture transcriptions can be segmented and assigned to their related slide with a high degree of reliability [130]. In this way, the textual content of the slides acts as metadata annotating the lecture stream. The text derived from slides is particularly valuable since it is likely to contain concise statements of the key points to be raised in the lecture [165] and to use carefully selected vocabulary specific to the topic under discussion. By contrast, the ASR transcription of the lecture will contain errors, and words that fall outside of the vocabulary of the recognizer will be missing entirely [86]. Social tagging has also been investigated in the lecture domain, and has been shown to improve access to lectures [140]. Students tagged lectures with handwritten information and photos, which not only added information, but served to reveal which parts of the lecture were most watched and potentially most interesting.

Podcasts are published on the internet with metadata at both the episode and the series level. The quality and completeness of the metadata varies from podcaster to podcaster, but it generally includes titles and descriptions. Internet search engines generally use metadata alone to index podcasts. Podcasts are an example of a case in which metadata provides information that is different from that which can be expected to be contained in the spoken content. When the user information need involves, for example, a speaker identity (cf. Figure 5.3), this information is more likely to be provided by the metadata than spoken by the speaker. Next we turn to ways in which useful metadata can be acquired or generated automatically.

5.2.3 Exploiting Automatically Generated Metadata

We have seen that metadata are useful for SCR not only to compensate for error in ASR transcripts, but also to provide additional information about spoken content that might correspond to dimensions of the user information need, but be unlikely to be contained in the transcripts. Here we overview the types of metadata that can be automatically extracted from speech media in order to aid SCR.

Speaker identification methods automatically match spoken content with the name of the person speaking. Much research has been devoted to developing methods to identify speakers independently of the words spoken [146]. Audio analysis can identify such non-verbal features as silence, music or speech, which are included, for example, in the inventory used in early work carried out by [314]. Other useful descriptors for speech media include speaker gender, channel characteristics (telephone vs. desktop microphone), speaker language, multiple speakers speaking at once, applause and coughing. Often, detection of these features is carried out as part of the process of speaker diarization, which will be discussed in more detail shortly. Automatic methods can also detect disfluencies and other non-lexical vocalizations [169]. Information about the location and frequency of disfluencies has the potential to provide clues as to the nature of speech, for example the formality of the style.

Further, affective information can be derived from spoken content, which allows the emotional dimension of speech media to be labeled [129]. An overview of technology for extracting emotional content from the speech signal is provided in [69]. Non-speech aspects of multimedia signals can also be analyzed in order to extract affective information [99]. In general, affective features can support SCR because of their correlation with particularly interesting or important segments of the spoken audio. Hot spots in meetings [307] are a specific example of where affect can serve to support navigation. In [234], the sound of baseball hits was combined with detection of excited speech in order to implement a system for highlights extraction from baseball games. Automatic laughter detection makes an important contribution to indexing spoken content in domains such as television sitcoms [123].

Audio event detection moves beyond what is produced by humans into the general domain of sound. Detectors can be built for identifying audio events [313]. In [219], experiments are performed using ten categories: *airplane jet*, *airplane propellor*, *birds*, *bus*, *cat meowing*, *crowd applause*, *dog barking*, *gun-shot*, *helicopter*, *horse walking*, *sirens*, *telephone bell*, *telephone digital*, *traffic*, and *water*.

Simple features sometimes reveal themselves as surprisingly effective. In particular, [63] highlights the usefulness of “surface features” — characteristics that are easily accessible to an indexing system, such as the length of an audio recording.

Automatic methods are useful for generating metadata, but they can also aid in simply collecting it. Meetings are an important domain for SCR, but one for which manual metadata are less readily available. However, collections of human-generated documents often exist within enterprises and contain helpful material specific to the meeting’s content. Research at IBM made an early contribution to the exploration of the automated delivery of information associated with a meeting [25]. The *Meeting Miner* system performs live ASR on the audio stream emerging from a meeting. The transcripts are used as a source of queries, which are extracted automatically and submitted to archives related to the meeting. The items returned are provided to participants to enhance their participation in the meeting. Information gathered in this way might potentially be used to annotate the meeting transcription or to more fully describe the topic under discussion in the meeting and thus potentially facilitate improved search. The key question here is whether materials can be chosen with sufficient selectivity and reliability to provide useful information to people involved in the meetings or interested in their content [143].

5.2.4 Expansion Techniques

This subsection examines techniques that can be used for expansions that make possible a better match between user queries and speech transcripts. Expansion techniques are drawn from text IR, where their benefits derive from their ability to provide lexical enrichment that compensates for semantic underspecification. In SCR, applying expansion techniques has the additional benefit of compensating for ASR errors. This subsection reviews techniques for the expansion of both queries and documents.

Query expansion. Expansion of queries can be accomplished with a variety of approaches. In subsection 4.4.3, we discussed the query expansion method introduced by [173, 174], which uses the language model and pronunciation dictionary to determine possible misrecognitions of the query word and uses these to expand the query. The method aims at compensating for speech recognition error.

Other query expansion methods have the effect of compensating for both ASR error and semantic underspecification. Here, we return to discuss relevance feedback, an IR technique initially introduced in subsection 2.2. In a standard relevance feedback scenario, users perform an initial search after which they provide feedback to the system indicating which retrieved items they deem relevant to their information need. In a pseudo-relevance feedback (PRF) scenario, top ranked items returned by an initial retrieval round are used as feedback, under the assumption that since they matched the query well they must be relevant. This feedback information is then used to modify the system to bias subsequent searches towards the information need.

PRF exploits co-occurrence of words within items to expand queries. Two mechanisms serve to make clear why relevance feedback has an overall tendency to reduce rather than amplify the effects of ASR error. First, recall that the match with a user's query is dependent on the presence of terms in the ASR transcript of an item. For spoken content items transcribed with low word error rates (WERs), there will be little impact in matching; items with higher WERs will be more significantly affected. We can then expect that items with lower WERs will appear at higher ranks in the results list, which was indeed observed by [237].

Because PRF chooses top-ranked items, it prefers well-recognized items to less well-recognized items. In short, PRF can be expected to have the tendency to disfavor items with worse WERs and the greatest chances of introducing word errors into the query.

Second, recall that the fixed vocabulary of an ASR system means that there are no spelling errors in the ASR transcript and no introduction of new words. For this reason, speech recognition transcripts do not contain very rare words. Rare words can be dangerous for relevance feedback, because they are highly specific and have large negative impact should they be inappropriately selected to expand a query. In [154], it is demonstrated that PRF can yield a greater percentage improvement in SCR tasks than in text IR tasks.

In work by [136, 137], a range of query expansion techniques making use of language resources (i.e., WordNet[1]), a collateral corpus and blind relevance feedback are explored. A recent approach to query expansion using a parallel corpus is presented by [189]. This approach uses topics discovered by way of dimensionality reduction in order to enrich user queries.

Document expansion. The process of query expansion extends queries using terms that have a strong statistical association with already-identified relevant items. These terms represent words that the user might have included in the original search request, but, for a variety of reasons, did not. By analogy we can consider the possibility of document expansion. We could use material relevant to a particular document as a source of terms to extend the document, with the goal of improving its representation of the underlying topic. In the case of spoken content, we are particularly interested in compensating for words missing from the ASR transcripts due to recognition errors, including OOV errors. This consideration motivated [251] to introduce document expansion for SCR. A selected document is used as a search request to a collection of text documents, and PRF methods are applied to select expansion terms for addition to the document transcript. Since there is no way of knowing whether the word has actually been spoken

[1] http://wordnet.princeton.edu/

or not, the technique strives to add words that were either actually spoken or that the user could have potentially spoken within the context of the document. Results on the TREC-7 SDR tasks by AT&T showed promise for this technique [250]. However, it has not been widely explored since this early work. In [192], phone confusion probabilities were used to expand documents, but this technique does not target the benefits of semantic enrichment. In particular, the relative benefits of document expansion for collections containing informal versus planned speech have yet to be investigated thoroughly.

Using collateral information. Collateral information is supplemental information derived from sources beyond the immediate collection of speech media. Collateral information can be used at many different stages in an SCR system. For example, in subsection 3.2.1, we mentioned its usefulness for adaptation of language models. Here, we turn to its usefulness for improving IR and for organizing and enriching speech media for the purpose of presentation to the user.

In comparison to text retrieval collections, speech media collections are typically relatively small. For retrieval, this means that parameters in the IR model could be poorly estimated, particularly with respect to the specificity of terms in individual items. The errors in ASR transcripts are likely to introduce further degradation of these estimates. This observation was made quite early in the development of SCR methods. In an attempt to address these problems, supplemental text corpora — much larger document collections, free of ASR errors — were successfully applied as a source of pseudo-relevant documents for PRF [137].

It should be noted that this technique is only effective if the collateral text corpora used are properly representative of the domain of the speech media collection. The SDR track at the TREC-7 and TREC-8 workshops used collateral text collections with notable success [127]. The TREC SDR materials were taken from North American radio and television news during a period in 1998. The text data sets that were used to augment retrieval from this collection consisted of text news stories from the same period. For domains that change rapidly over time, it is important that the data is not just in the same topical space, but that it is from the same time period since important items of the vocabulary and their usage will often be significantly different, even from those in the previous or following year [134].

Another important source of collateral information is closed captions, which are generally more accurate than contemporary ASR systems and can be used to support SCR [103, 206]. In such situations the only reason to perform ASR would be to obtain an exact alignment between the spoken content and the transcript. Forced alignment techniques are treated in more detail in the following discussion.

5.2.5 Forced Alignment

Forced alignment is the process of using the ASR system to temporally match speech media with related material, usually with a human-generated transcript. The alignment is "forced" in the sense that the ASR system is constrained so that it recognizes only the words contained in the transcript and only in the sequence in which they occur in the transcript. Instead of recognizing the words themselves, the ASR system is recognizing the positions of words, whose identity has been supplied in advance. The result is an enrichment of the original transcript in which each word is associated with a time code that indicates where that word was spoken in the speech stream. An example of the use of forced alignment is the *Radio Oranje* system [109], in which human-generated transcripts of speeches of the Queen of the Netherlands are synchronized with historical recordings. After alignment, the transcripts can be indexed and used to provide search functionality for users. Users can carry out searches and the system returns results that correspond to jump-in points in the speech stream. We discuss the usefulness of forced alignment for user interfaces further in *Accessing Information in Spoken Content*. An overview of uses for alignment techniques is available in [62].

In [171], forced alignment is performed between stenographic transcriptions of parliamentary speeches and recordings made in the parliament. The match between these two resources is close, but inexact. In order to carry out forced alignment, the transcriptions are used to generate lattices that allow for the deletion and skipping of words.

These lattices are then matched to the speech signal in order to create a correspondence between the spoken word and the transcribed content.

Forced alignment can be considered a method for expansion, since it associates speech media with additional information from an aligned transcript. However, it can also be considered a structuring method. Any structure that exists within the transcript, for example, paragraph markings, defines a structural spoken content unit when it is aligned with the speech media. In the next subsection, we turn to additional techniques for structuring spoken content.

5.3 Structuring and Representing Spoken Content

Segmentation techniques and techniques for generating alternate semantic representations create information about spoken content that is useful for IR. Here we introduce techniques and comment on their interaction with IR algorithms. In *Accessing Information in Spoken Content*, we further discuss how these techniques can be used in the user interface.

5.3.1 Segmentation

Segmentation of spoken content can take place either based on direct analysis of the audio content or by using the ASR transcripts. The importance of forming audio and topical segments has long been recognized in management of speech content. Early work on this subject was conducted on topic and speaker identification using HMM-based speech recognition in studies such as [85]. Segmentation using the audio content prior to recognition can help to improve the quality of the ASR transcripts. In general, ASR systems do not process a continuous string of speech, but rather disassemble it into smaller pieces, generating a hypothesis string for each. Segments created for the purposes of ASR may be of fixed length or may be divided at pauses (i.e., places at which the speech signal drops off in energy). Segments may roughly correspond to utterances, but whether they will also be useful for SCR will depend on the application. Some recognizers hypothesize punctuation and output units that can be equated with sentences [18, 169]. Such units are more clearly directly semantic in nature and helpful for SCR. In general, the quality of the segmentation will be strongly dependent on the segmentation algorithm used and on the nature of the audio signal being segmented.

Multiple levels of segmentation may be relevant for an SCR system. For example, in an interview, both topical segments and individual speaker turns may be relevant. The IR model may need to combine the scores. Score combination is not trivial: a larger item may score well on its own because overall it contains a large number of terms, but it may still not specifically address the information need of the particular query. In contrast, individual segments may score well without being representative of the document as a whole. One example of an approach from text IR that is potentially helpful in SCR because it combines different levels of scores is described in [30]. Here, the whole document score is combined with the highest scoring segment from within the document. The appropriateness of this approach will depend on the application. Some techniques carry out topic labeling and segmentation simultaneously, and we will return to discuss such approaches in subsection 5.3.4.

5.3.2 Diarization

Much work on structuring spoken content has taken place within the context of the research effort devoted to *diarization* systems. Diarization is the task of automatically identifying sections of spoken audio and correctly labeling them with their characteristics, for example, speech, non-speech, male-speech, female-speech, music, noise. Although speaker identification played a role in early segmentation approaches, e.g., [300], determination of the identity of the speaker, called *speaker identification*, or confirmation of a presumed speaker identity, called *speaker verification*, does not fall into the scope of the diarization task.

Work on diarization was promoted by the Rich Transcription (RT) Evaluation Project [74, 317][2] of the US National Institute of Standards and Technology, which ran from 2002–2009. Rich transcripts are ASR transcripts that include information above and beyond conventional ASR transcripts. The enriching information makes the transcripts more

[2] http://www.itl.nist.gov/iad/mig/tests/rt/

readable for humans and more useful to machines. The RT Evaluation initially used English-language broadcast news, telephone speech and meeting room speech. Later, it ran tasks involving Mandarin and Arabic. Tasks included metadata extraction (MDE) tasks such as detection of words used by speakers as fillers or to correct themselves (i.e., filler and edit word detection), speaker attributed speech-to-text (transcribe the spoken words associated with a speaker) and speech-to-text (STT) tasks such as speaker diarization.

Diarization techniques usually make use of a technique based on the Bayesian Information Criterion (BIC). A sliding window is moved along the speech signal, which is represented as a series of "speech vectors" encoding information about signal properties such as frequency and energy. For each window position, the BIC is used to decide whether a single model or two different underlying models provide a better explanation (i.e., "fit") for the observed speech signal. If a window is better explained by two underlying models, then a segment boundary is hypothesized to bisect that window. Once segment boundaries have been located in this manner, clustering is carried out in order to determine which segments are similar to each other, that is, were spoken by the same speakers. An audio classifier, trained on labeled training data, is then used to assign a label to each group of segments. Early work on diarization includes [44, 95]. Subsequently, much research has been dedicated to speaker diarization, with an overview available in [266] and recent work including [6, 118, 309, 316]. The diarization task can be extended to include automatic methods for discovering the role that speakers play within conversations, such as investigated by [274].

5.3.3 Compact Representations

Techniques that capture the essence of spoken content and represent it in succinct form can be used both for indexing purposes and for the purposes of visualizing speech media, which is relevant to the discussion of SCR interfaces in Section 6. Here we cover several techniques that go beyond ASR in that they represent spoken content in a manner more condensed than a word-for-word transcription.

Automatic term extraction. Term extraction can be viewed as a specialized summarization technique. Automatic term extraction can be used for the purpose of providing a select set of indexing terms for SCR. Frequently, however, automatically extracted terms are used to represent speech media for the users in the interface. In [65], term extraction was investigated for the broadcast news domain. The technique was shown to produce useful keyphrases at relatively high word error rates. In [101], terms are extracted from highly noisy transcripts and ranked in terms of descriptiveness. The usefulness of extracted terms displayed to the user in the form of a term cloud is investigated in [268]. A term cloud is a collection of terms that is visually weighted, in other words, the size and bolding of the font in which a term is displayed reflects its importance. Importance is usually calculated on the basis of term counts within an item. When used as a representation of spoken content, a term cloud is created using term frequencies of words occurring in the speech recognition transcripts. Because a word needs to occur multiple times in the transcripts in order to be included in the cloud, term clouds have the potential to compensate for ASR errors. The largest, most important terms in the cloud are highly likely to have actually occurred in the spoken content and not to be the result of ASR errors.

The Cross-Language Speech Retrieval track (CL-SR) [205, 217, 295] carried out SCR experiments that compared the usefulness of automatically-assigned keywords versus manually-assigned keywords. The results clearly indicate that much work must be done before automatically generated metadata can rival human generated metadata in usefulness.

Summarization. Summarization of spoken content takes two basic forms: speech-to-text summarization and speech-to-speech summarization [79]. If video is also taken into account, then both static storyboard-type summaries as well as video clip summaries can be considered. The two basic approaches to speech summarization, as identified by [79], are sentence extraction — in which entire sentences are selected from the spoken content and then concatenated — and sentence compaction — in which sentences are shortened or the transcript is modified. As with SCR, the main differences between speech and text summarization arise due to ASR errors and also due to the unstructured nature of spoken content. However, speech also has prosodic cues, which can help to create summaries. For example, in [42], emphasis is used to help automatically summarize spoken discourse. Recent work on summarization includes research in the area of lectures and presentations [80], voicemail summarization [149], and broadcast news summarization [45].

Another interesting type of summarization task is title generation. For example, in [126], a system is presented that uses speech recognition transcripts to automatically generate titles for a collection of news stories. Automatic title generation, as well as automatic summaries, is provided by the Chinese news prototype described in [166]. The automatic generation of lecture slides from ASR transcripts is explored in [130].

5.3.4 Extracting Topic Information

In its effort to match user queries to collection items, an SCR system needs to deal successfully with the problem of *vocabulary mismatch*. Vocabulary-mismatch arises because, in human language, it is possible to discuss one topic using a wide variety of lexical items. Vocabulary-mismatch poses a challenge for both text IR and SCR, although in the SCR case it can be more extreme due to ASR error. An approach to dealing with mismatch is to choose abstract representations of topics and then attempt to assign specific spoken content items to one of these abstract representations. Topic inventories are either pre-determined and take the form of a categorization scheme or an ontology, or else they can be determined dynamically by analyzing the content. In the remainder of this section, we discuss techniques that belong to both approaches.

Spoken content classification. Category schemes group similar items, both for the purposes of computing similarity within the SCR system and for display to the user. The widespread use of category schemes to organize spoken content attests to their ability to capture useful semantic regularities among spoken media items. For example, users looking for podcasts on iTunes can make use of topic categories in their searches. Category labels include, *Arts*, *Science and Medicine*, and *Technology*.[3] Each of these larger labels can be broken down into sub-labels.

The task of spoken audio classification involves assigning items to categories. Generally, a classifier is trained on a set of labeled training examples and then used to assign these labels to new, yet unseen, items. The collection is usually considered to be static. The first work done in the area of automatic classification of spoken audio was likely [231], which describes a system for classifying spontaneous speech messages into one of six classes related to topic. Text classification techniques can be applied to speech transcripts in order to automatically tag spoken audio with subject tags. One of the major goals of automatic classification research is to reproduce the classification that would be generated by a human, given a specific classification scheme. For example, in [213], classes are drawn from the *Media Topic Taxonomy* of the International Press Telecommunications Council,[4] in [159] classes are drawn from the thesaurus used by archivists at the Netherlands Institute for Sound and Vision, a large audio–video archive, and in [161] classes are drawn from the set of tags assigned by users to video on blip.tv, a video sharing platform. Note that some forms of classification that are potentially useful for SCR perform their categorization of speech media based on characteristics not directly related to topic, such as genre in video [267] or dialogue acts in spoken conversation (i.e., statement, question, apology) [259].

Latent topic analysis. The vocabulary mismatch problem can also be addressed with latent topical analysis, a process that maps word forms onto dimensions of meaning. These dimensions are referred to as "latent" because they are expressed in terms of word co-occurences, but are not otherwise explicitly present in the content. In contrast to topic category schemes, the dimensions are not known in advance and may not always be naturally interpretable to humans, even though

[3] http://www.apple.com/itunes/podcasts/specs.html#categories

[4] http:www.iptc.org

they do succeed in capturing useful underlying semantic regularities. The technique of Latent Semantic Indexing (LSI) was introduced for IR in [64]. LSI involves applying Singular Value Decomposition (SVD) to a matrix containing information about which terms are included in which documents. The SVD technique identifies latent dimensions in their order of importance and top dimensions are then used as indexing terms. Words from the documents and queries are mapped onto these new indexing terms, meaning that retrieval takes places within the latent semantic space. Work in the area of SCR that has exploited LSI and related techniques includes [40, 115, 152].

Topic segmentation. We close this subsection with a discussion of techniques that consider structure and topic not as two separate steps, but rather as part of an integrated process. Speech media can be partitioned into topic-homogenous segments by using methods adopted from text segmentation. A popular method for automatically determining topic boundaries is the TextTiling algorithm [106], which hypothesizes topic boundaries at points at which a major shift in the vocabulary use is observed. Alternative segmentation algorithms are described in [48, 66] and [175]. The work in [17, 116] concentrates specifically on segmentation of meetings and dialogues. TextTiling operates on words alone, but other characteristics of the speech signal can also be exploited for topic segmentation. For example, in [260], lexical items and prosody are combined in order to predict topical segments.

Topic detection and tracking (TDT). Topic detection and tracking (TDT) is a suite of tasks aimed at both segmenting a speech stream and identifying the topics that it contains. Much productive research effort was devoted to these tasks during the TDT program [287],[5] which was sponsored by the US Department of Defense during the years 1998–2004. TDT conducted tasks for topic segmentation, tracking and detection as well as tasks for detecting first occurrence of new topics and also for linking items that are topically related. TDT focuses on the discovery of new material and emphasizes events over thematic

[5] http://www.itl.nist.gov/iad/mig/tests/tdt/

subject categories [4]. The TDT research program comprises five subtasks: *Story Segmentation*, detection of boundaries between stories in a news show, *First Story Detection*, detection of new, unknown stories, *Cluster Detection*, grouping incoming stories into topically related groups, *Tracking*, monitoring the incoming stream to find more stories resembling a set of sample stories, and *Link Detection*, for deciding whether a given pair of stories is related or not. Detailed descriptions of tasks and performance can be found in [5].

TDT techniques were developed for deployment in a system that monitors the speech media stream and generates an alert when interesting items are encountered. Ultimately, all techniques covered in this section that move beyond basic ASR transcripts are only useful in so far as they can support SCR and users of SCR systems.

This section has covered methods for moving beyond ASR for the augmentation, structuring and representation of speech media. It concludes our treatment of the component technologies for SCR and techniques for combining them. In the next section, we turn to the issue of the interfaces that support user/system interaction, which play a vital role in satisfying users' needs.

6

Accessing Information in Spoken Content

The ultimate usefulness of a Spoken Content Retrieval (SCR) system to users is determined by the user interface. Technically, an SCR system may return results representing high quality matches with the query. However, users must be able to efficiently evaluate these results and identify individual items of interest in order for an SCR system to fulfill its function of satisfying user information needs. Further, the interface should make full use of feedback from users in order to refine queries. As mentioned in the *Introduction*, SCR systems have conventionally paid little attention to user requirements. The interface has been explicitly identified as an important SCR challenge in [11, 203], and representation of spoken content is noted as an important open issue for SCR in [3]. This section overviews the key issues of interface design and discusses classic examples of how these issues have been addressed in research prototypes and real-world systems. We focus on aspects that are specific to SCR, with the intent of providing a helpful complement to the existing literature on user interfaces for search, such as [107].

6.1 Query Entry

SCR systems often invite users to enter queries by offering a query box, such as is in widespread use with Web search engines. The query interface of PodCastle (http://podcastle.jp) [92, 93], a browser-based SCR system for podcasts, is shown in Figure 6.1 to illustrate such a query box. Although the simplicity of the query box echoes the Google Web search engine (http://www.google.com), PodCastle offers the user much more information concerning the search system and what it can be used to find. For example, the scope of the content indexed (over 100,000 episodes) is explicitly mentioned and links to recommended podcast episodes are provided.

The Google election gadget [2], which offered search functionality for political speeches during the 2008 US elections, also displayed hints for users. It provided information about the scope and the functionality of the system by prompting users with the question "What did the

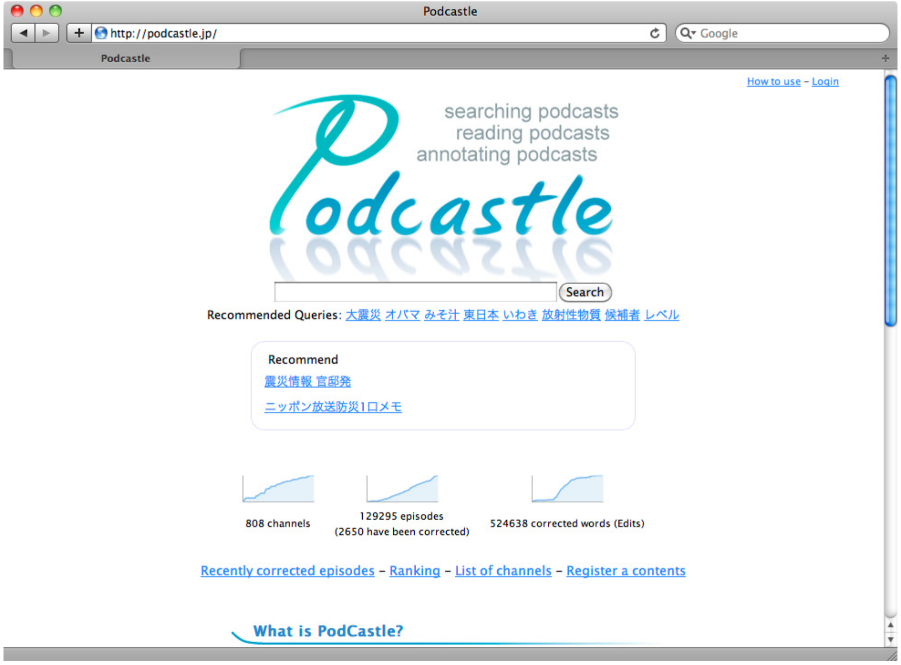
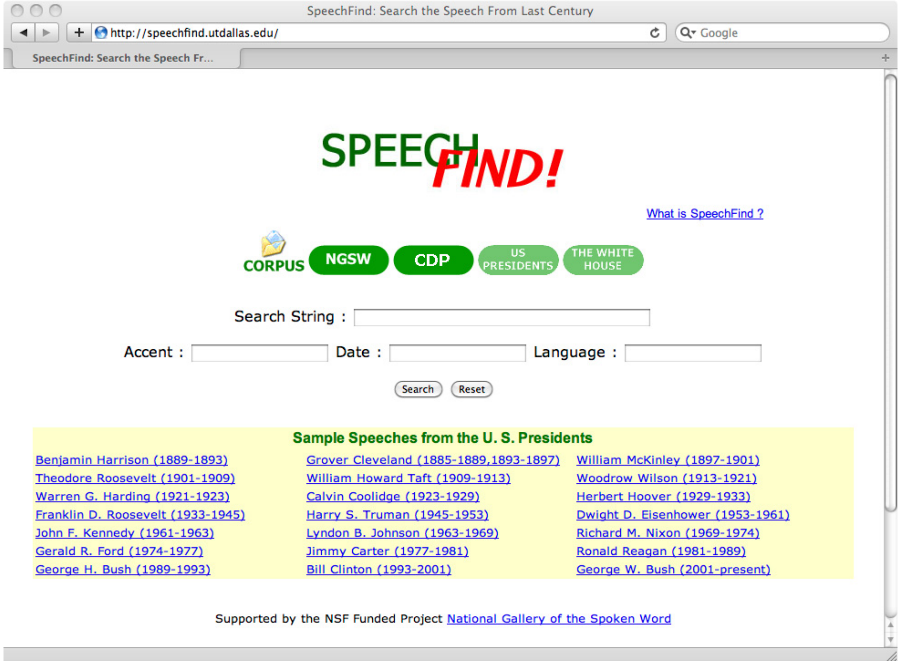


Fig. 6.1 Query interface of Podcastle.

candidates say?" In addition to the query box, there was a choice of constraining the search to "McCain," "Obama" and "Debates." These choices implicitly supply information about what is present in the system. If users limit the field of search to a particular category of content, the accuracy and speed of the system stands to improve. In general, the trend can be observed towards designing query interfaces that inform and guide the user. Because users are not yet widely aware of the existence of spoken-content-based retrieval technology, as suggested by the user study in [21], this information may support them in formulating more effective queries.

SpeechFind [100, 144] is another system whose query interface offers users information about what can be found in the system. As shown in Figure 6.2, SpeechFind displays buttons at the top that show the specific sources of content available in the system and also displays sample content below.



Sample Speeches from the U. S. Presidents

Benjamin Harrison (1889-1893)	Grover Cleveland (1885-1889,1893-1897)	William McKinley (1897-1901)
Theodore Roosevelt (1901-1909)	William Howard Taft (1909-1913)	Woodrow Wilson (1913-1921)
Warren G. Harding (1921-1923)	Calvin Coolidge (1923-1929)	Herbert Hoover (1929-1933)
Franklin D. Roosevelt (1933-1945)	Harry S. Truman (1945-1953)	Dwight D. Eisenhower (1953-1961)
John F. Kennedy (1961-1963)	Lyndon B. Johnson (1963-1969)	Richard M. Nixon (1969-1974)
Gerald R. Ford (1974-1977)	Jimmy Carter (1977-1981)	Ronald Reagan (1981-1989)
George H. Bush (1989-1993)	Bill Clinton (1993-2001)	George W. Bush (2001-present)

Supported by the NSF Funded Project National Gallery of the Spoken Word

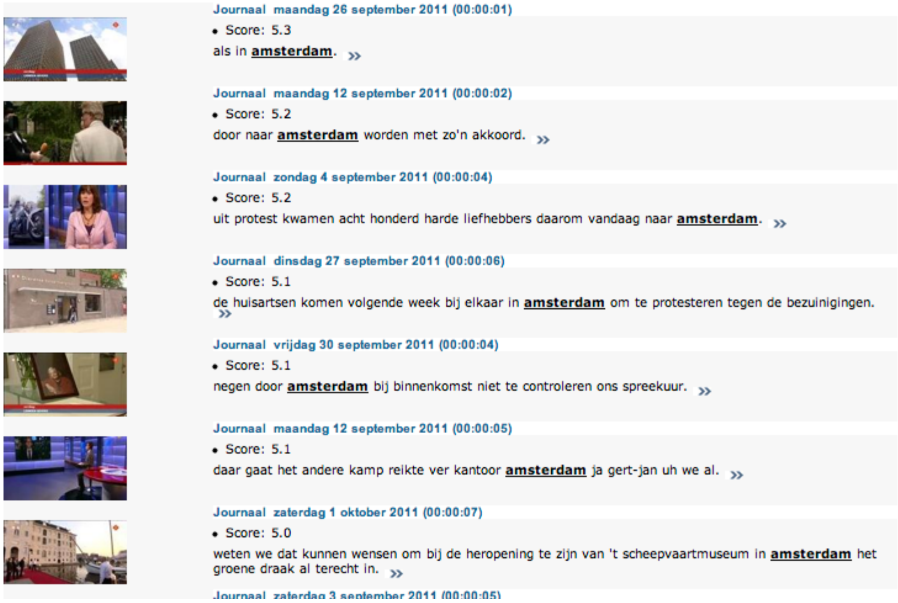
Fig. 6.2 Query interface of Speechfind.

The observation that longer, richer queries lead to better SCR performance has led to the suggestion that interfaces consider how users can be encouraged to formulate longer queries [3]. The additional boxes for query entry, as depicted in Figure 6.2, both invite the user to enter information of a specific type and help the system to disambiguate between query terms that the user expects to hear spoken in the audio content and other characteristics of the speech media, such as the identity of the speaker. This form of query entry resolves ambiguity in queries containing person names, as discussed by [21] and mentioned in subsection 5.1. Ultimately, users will probably find entering a structured query too cumbersome and systems that automatically differentiate the relationship between query terms and different components of the user information need will prove more effective.

As a final comment on queries, we mention that in addition to text queries, SCR systems can also enable users to specify non-textual queries to initiate the search process. For example, the Informedia system offers functionality, described in [278], that allows a user to initiate a query by specifying a location on a map. The system returns news stories containing references to places in this region. Other functionality, such as described in [49], makes it possible to query by visual features such as visual concept or image. Finally, we mention again that voice search, which uses spoken queries, is a rapidly growing area [285].

6.2 Display of Results for Selection

Results returned by the SCR system are generally displayed as a results list. As a guiding principle, selection interfaces of SCR systems should be designed to let users make maximum use of their innate human ability to quickly ascertain the interest and relevance of particular objects [202, 203]. The importance of informative result display is reflected in, for example, [265], which presents the results of a user study suggesting that multimedia retrieval may be affected if the user has a low perception of the relevance of the retrieved results. The results display should be optimized in order to allow quick and easy assessment of relevance by users.



Journaal maandag 26 september 2011 (00:00:01)

- Score: 5.3

als in **amsterdam**. »

Journaal maandag 12 september 2011 (00:00:02)

- Score: 5.2

door naar **amsterdam** worden met zo'n akkoord. »

Journaal zondag 4 september 2011 (00:00:04)

- Score: 5.2

uit protest kwamen acht honderd harde liefhebbers daarom vandaag naar **amsterdam**. »

Journaal dinsdag 27 september 2011 (00:00:06)

- Score: 5.1

de huisartsen komen volgende week bij elkaar in **amsterdam** om te protesteren tegen de bezuinigingen. »

Journaal vrijdag 30 september 2011 (00:00:04)

- Score: 5.1

negen door **amsterdam** bij binnenkomst niet te controleren ons spreekuur. »

Journaal maandag 12 september 2011 (00:00:05)

- Score: 5.1

daar gaat het andere kamp reikte ver kantoor **amsterdam** ja gert-jan uh we al. »

Journaal zaterdag 1 oktober 2011 (00:00:07)

- Score: 5.0

weten we dat kunnen wensen om bij de heropening te zijn van 't scheepvaartmuseum in **amsterdam** het groene draak al terecht in. »

Journaal zaterdag 3 september 2011 (00:00:05)

Fig. 6.3 Excerpt of results list generated by the Dutch Broadcast News Retrieval system of the University of Twente in response to the query "Amsterdam."

We use an excerpt from a results list generated by the Broadcast News Retrieval system of the University of Twente [211] in Figure 6.3 to illustrate some key issues of results display. The results list has been returned by the system in response to the query "Amsterdam." The results are displayed as a ranked list of jump-in points each associated with a speech media fragment that the system has matched with the query. Each item is represented by a surrogate comprising a short excerpt of the ASR transcript, a keyframe, the name and date of the program and a timecode locating the result within the program. The function of the surrogate is to give the user the information necessary to evaluate the relevance of the result to the original information need and to decide whether to review the result in more depth. Surrogates are also used in text retrieval, but are particularly important for SCR systems. Reviewing a spoken media result requires listening to an audio file or watching a video file. This process is considerably more time consuming that skimming a page of text. The benefit of snippets is related to the quality of the underlying speech transcripts — in [108] it is observed that the accuracy of surrogates determines their usefulness. If subword indexing is used, subword units need to be reconstituted into words before they are appropriate for use in a snippet.

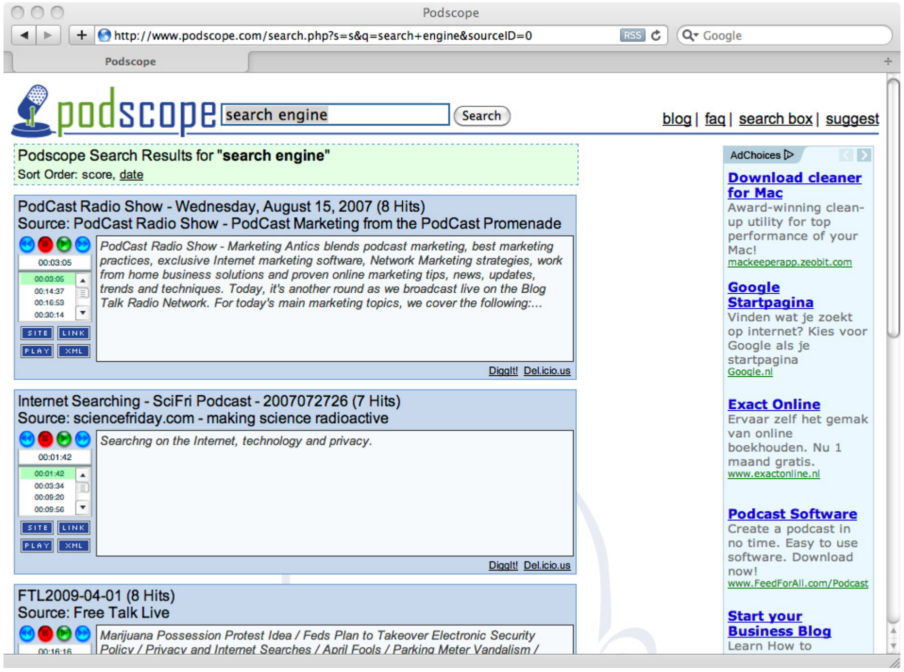
The surrogate is usually "biased" towards the user query, meaning that its form is especially chosen to highlight the match between the query and the result. Note that in Figure 6.3, the query word has been highlighted in each snippet. The presence of the query word is strong evidence for the user that the result is relevant to the information request. Interface design should take into consideration the user's expectation level of seeing the query word in the snippet and hearing the query word very quickly after the playback of the result is initiated. For video applications, presenting a keyframe may provide the user with an additional hint as to the content of the result. A further dimension to the selection of the appropriate form of surrogates is the background knowledge of the user, which has been observed to have an impact on the types of surrogates that are preferred [158, 268].

Additionally, it is important to mention how results selection display may effectively limit the application of advanced IR techniques in practice. In *Spoken Content Retrieval beyond ASR Transcripts*, IR techniques that are capable of overcoming words missing in the transcripts due to speech recognition errors were discussed, in particular, query expansion. However, such techniques may return result items that are relevant to the user's original information need, but where none of the original query words are actually uttered in the spoken content. Unless there is a mechanism to convince the user of the relevance of a result without showing evidence of the query word being directly associated with the content, users will pass over this result and the system may fail to meet its goal of satisfying the user information need.

One of the challenges of displaying a ranked list of results is to effectively communicate to the user the relationship between the results and the structure of the speech media in the underlying collection. Recall, from Section 5, that multiple levels of units may be used by the SCR system and that the retrieval unit is not necessarily the only important or useful unit of structure within the collection. For example, in Figure 6.3 there is a tension between the retrieval unit (a fragment) and a larger, natural unit in the collection (a news item). Two results are returned from the same news program on Monday, 12 September 2011. Depending on the application, two results from the same program might confuse a user, who may consider them actually to constitute a single, duplicated result. Even if they contain different spoken content, it is difficult to indicate the difference clearly in the results list because, as illustrated by this example, results display often depends on program level metadata, in this case the date, which is the same for each.

A typical approach is to choose the larger unit with which metadata is associated as the retrieval unit that is ranked and displayed in the results list. This approach is taken by PodScope (http://www.podscope.com), a spoken-content-based podcast search engine. A results list returned by PodScope in response to the query "search engine" is displayed in Figure 6.4.

Each result is an individual podcast episode displayed together with its metadata. Note that information about the relevance of individual


Podscope
http://www.podscope.com/search.php?s=s&q=search+engine&sourceID=0
RSS
Google
Podscope
podscope
search engine
Search
blog | faq | search box | suggest
SITE
LINK
PLAY
XML
Diggit! Del.icio.us
AdChoices

Podscope Search Results for "**search engine**"
Sort Order: score, date

PodCast Radio Show - Wednesday, August 15, 2007 (8 Hits)
Source: PodCast Radio Show - PodCast Marketing from the PodCast Promenade

00:03:05
00:03:05
00:14:37
00:16:53
00:30:14

PodCast Radio Show - Marketing Antics blends podcast marketing, best marketing practices, exclusive Internet marketing software, Network Marketing strategies, work from home business solutions and proven online marketing tips, news, updates, trends and techniques. Today, it's another round as we broadcast live on the Blog Talk Radio Network. For today's main marketing topics, we cover the following:...

Internet Searching - SciFri Podcast - 2007072726 (7 Hits)
Source: sciencefriday.com - making science radioactive

00:01:42
00:01:42
00:03:34
00:09:20
00:09:56

Searchng on the Internet, technology and privacy.

FTL2009-04-01 (8 Hits)
Source: Free Talk Live

00:16:16

Marijuana Possession Protest Idea / Feds Plan to Takeover Electronic Security Policy / Privacy and Internet Searches / April Fools / Parking Meter Vandalism /

Download cleaner for Mac
Award-winning clean-up utility for top performance of your Mac!
mackeeperapp.zeobit.com

Google Startpagina
Vinden wat je zoekt op internet? Kies voor Google als je startpagina
Google.nl

Exact Online
Ervaar zelf het gemak van online boekhouden. Nu 1 maand gratis.
www.exactonline.nl

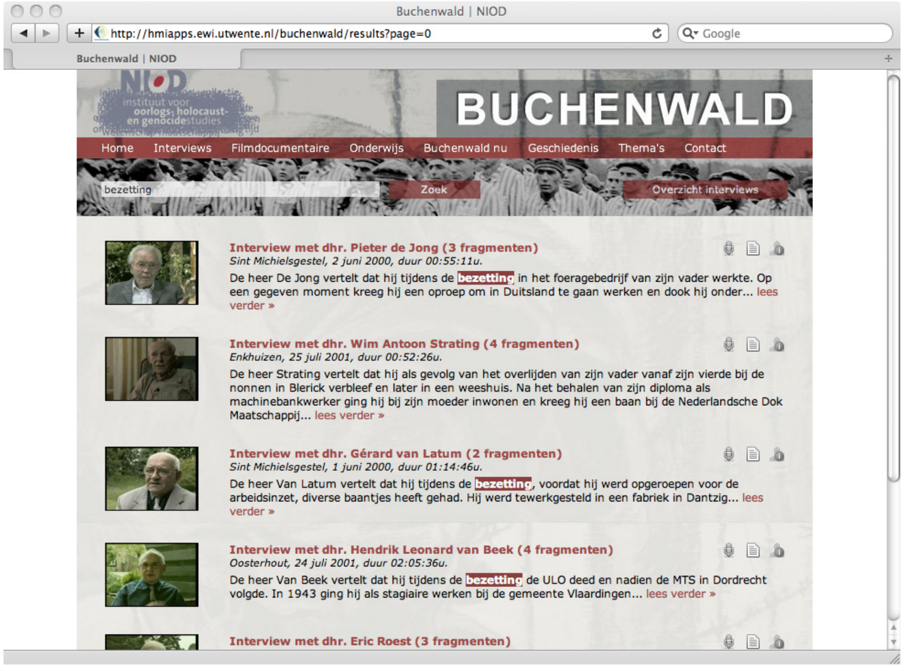
Podcast Software
Create a podcast in no time. Easy to use software. Download now!
www.FeedForAll.com/Podcast

Start your Business Blog
Learn How to

Fig. 6.4 Results of the Podscope search engine in response to the query "search engine."

fragments within the podcast episode is contained in the surrogate. On the left is a scrolling list of jump-in points, displayed as time codes, which allow the user to initiate playback of a particular fragment-level result directly from the episode-level results list. Displaying both episodes and fragments together in the results list gives greater flexibility for result review. However, this benefit is offset by the relatively large amount of space required by the surrogate and the fact that time codes provide little information to the user about which fragment would be most interesting to select. Another approach to results display is represented by the search application developed by the University of Twente for the collection at http://www.buchenwald.nl containing interviews with Dutch survivors of the Buchenwald concentration camp [211].

A results list returned in response to the query "bezetting" (Eng. "occupation") is displayed in Figure 6.5. Each result is an individual



Interview met dhr. Pieter de Jong (3 fragmenten)
Sint Michielsgestel, 2 juni 2000, duur 00:55:11u.
De heer De Jong vertelt dat hij tijdens de **bezetting** in het foeragebedrijf van zijn vader werkte. Op een gegeven moment kreeg hij een oproep om in Duitsland te gaan werken en dook hij onder... lees verder »

Interview met dhr. Wim Antoon Strating (4 fragmenten)
Enkhuizen, 25 juli 2001, duur 00:52:26u.
De heer Strating vertelt dat hij als gevolg van het overlijden van zijn vader vanaf zijn vierde bij de nonnen in Blerick verbleef en later in een weeshuis. Na het behalen van zijn diploma als machinebankwerker ging hij bij zijn moeder inwonen en kreeg hij een baan bij de Nederlandsche Dok Maatschappij... lees verder »

Interview met dhr. Gérard van Latum (2 fragmenten)
Sint Michielsgestel, 1 juni 2000, duur 01:14:46u.
De heer Van Latum vertelt dat hij tijdens de **bezetting**, voordat hij werd opgeroepen voor de arbeidsinzet, diverse baantjes heeft gehad. Hij werd tewerkgesteld in een fabriek in Dantzig... lees verder »

Interview met dhr. Hendrik Leonard van Beek (4 fragmenten)
Oosterhout, 24 juli 2001, duur 02:05:36u.
De heer Van Beek vertelt dat hij tijdens de **bezetting** de ULO deed en nadien de MTS in Dordrecht volgde. In 1943 ging hij als stagiaire werken bij de gemeente Vlaardingen... lees verder »

Interview met dhr. Eric Roest (3 fragmenten)

Fig. 6.5 Results of the search application for the Dutch language oral history interview collection at http://www.buchenwald.nl in response to the query "bezetting."

interview. The user is not presented with the relevant fragments directly, but rather merely supplied with information on how many there are (e.g., three fragments is indicated with "3 fragmenten"). The user enters an interview in order to explore the fragments that it contains. Here again, as already mentioned in subsection 5.3.1, segmentation interacts with IR models. In order to rank interviews, it is necessary to combine fragment-level relevance to an overall interview-level score. The exact balance to be used when making this computation is important. For example, in some cases a unit containing a single highly relevant fragment should out-rank a unit containing multiple less relevant fragments. In other cases, the opposite could be expected to hold. This combination itself should depend on how users use the system.

6.3 Review and Playback of Results

SCR systems include a player that allows the user to play back individual results. Playback is an important functionality for both content review and content consumption.

Playback should have the goal of saving the user time and reducing memory load, since humans generally have poor facilities for extracting and remembering detailed information from audio content [112]. As a design guideline, players should provide as much information as is possible, without clutter, about the content of speech media and give users flexible control over navigation. Naturally, an important part of accessing spoken content will remain actually listening to audio material. The process can be made more efficient by enabling fast playback. In fact, experiments show that speech can still be intelligible if the speed of delivery is doubled [9]. The SpeechSkimmer system used different levels of compression for fast playback [9, 10]. However, the cognitive load of listening to speech at this speed is considerably higher than natural speech, leading to the listener rapidly losing focus or becoming overloaded. In the end, the gain is also only half of realtime, which means that reviewing speech media results remains a time consuming process. Other approaches for compressing speech involve not only altering the speech rate but also removing unimportant words and segments (called *excision*) [269, 270].

Another option available to SCR systems is presenting users with ASR transcripts to read. Scanmail [297] is an example of a system that presents the user with the ASR transcript directly, in this case a voicemail message. A user study suggested that this feature was appreciated. However, a relationship does exist between the error levels of ASR transcripts and their usefulness in the system, with higher error rates being less useful [194, 258]. In [194], a user study on the usefulness and usability of ASR transcripts for a web archive was conducted. Transcripts with WER $> 45\%$ were found to be unsatisfactory while transcripts with WER $< 25\%$ were found to be useful and usable.

It is important to keep in mind, that an SCR system must not allow users to develop an unfounded trust in the ASR transcripts. In [35], professional users were found to have significant confidence in the SCR system, the transcripts and their own ability to work with them. This resulted in the users failing to seek relevant content not explicitly reflected by the transcripts, reducing the recall of their results. The same effect was reported in user tests of the Scanmail system [297]. In the domain of voicemail, recall is more critical and misplaced trust in the ASR transcripts caused users in the study to miss crucial information that was not recognized by the ASR system.

If ASR transcripts have word- or sentence-level time codes, these can be offered to users to read in the interface, linked to the speech stream so that users can click in the transcript to jump directly to listening to the stream, when they find a portion that interests them. Effectively, the text of the transcript becomes the playback interface. An example of an audio browser, dating back to the early 1980s, with an interface that linked text with speech is the *intelligent ear* [242]. This system presented a representation that depicted the amplitude of the waveform. During playback the current play position is indicated by a sound cursor that moves forward synchronously with playback and highlights that portion of the wave currently being played. Selected keywords spoken in the audio are written in under the waveform. Keywords that are recognized with higher confidence are displayed more brightly. Modern interfaces tend to present variations on this basic theme. A key feature is that the interfaces are linked to the media recording and that the user can initiate playback at any point by clicking on the corresponding point in the interface.

Typically SCR players combine functionality that allows users random access to the speech stream with a tape-recorder metaphor, unchanged since early mentions such as [29, 76]. In these interfaces, time runs from left to right [29], with events positioned along the timeline proportionately to where they occur in the broadcast. Thought should be devoted to making the time scale consistent in order to facilitate comparison [202]. Many interfaces incorporate a slider bar that the user can manipulate to control the point of playback, which doubles as a timeline. The timeline contains markers indicating the positions of particularly relevant material. An early example of this functionality is illustrated in the Video Mail Retrieval system browsing interface [27], shown in Figure 6.6. The interface shows a graphical timebar with

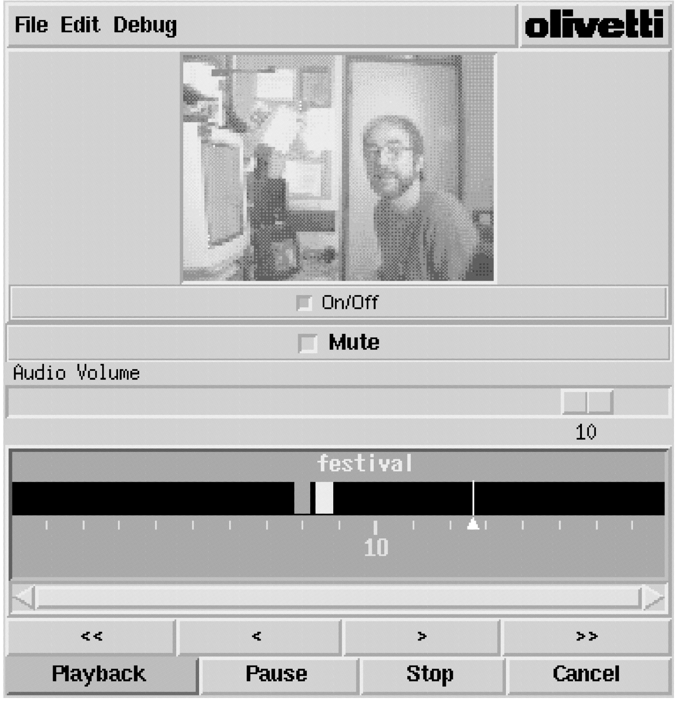
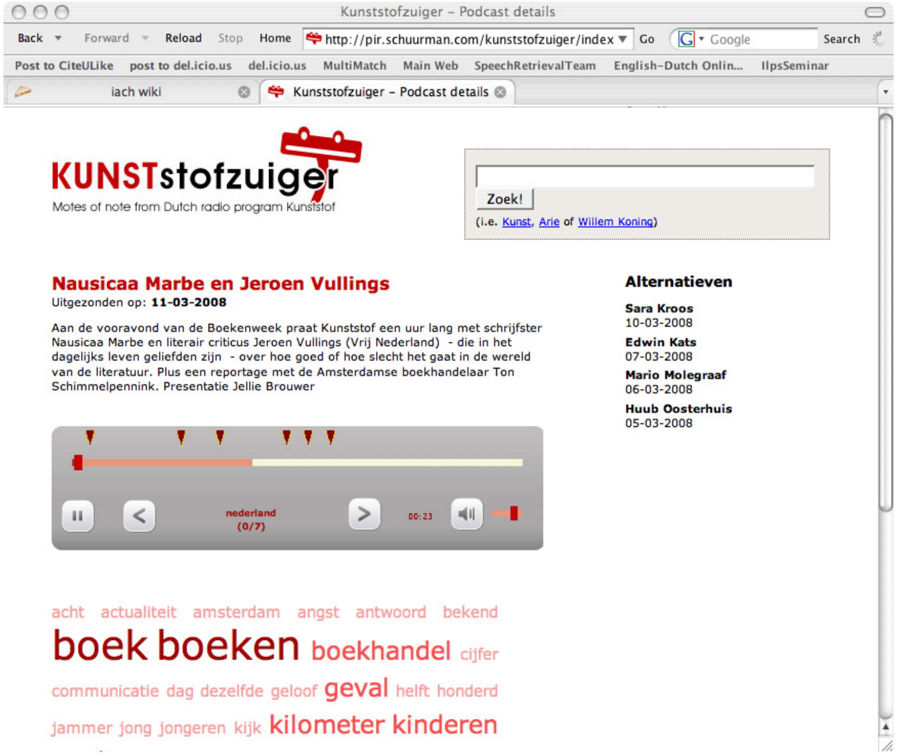


Fig. 6.6 Early Video Mail Restrieval system interface.

individual hits on query words in the audio file highlighted. Similar to [242], search term confidence is indicated by the brightness of the results when displayed. The user can click to start playback at any point on the timeline. This example is only 20 seconds in length — for longer files, clusters of search term hits can direct the user to regions more likely to be relevant to the query.

If the speech media is a video, a storyboard of clickable keyframes can be used to depict the temporal progression graphically for the user. An example of this principle in use is the CMU News-On-Demand system [105].

Figure 6.7 depicts a Dutch podcast search engine called *Kunststofzuiger* [46] developed at the University of Amsterdam, chosen to illustrate typical strategies for SCR players. The player page displays


Kunststofzuiger – Podcast details
Back
Forward
Reload
Stop
Home
http://pir.schuurman.com/kunststofzuiger/index
Go
Google
Search
Post to CiteULike
post to del.icio.us
del.icio.us
MultiMatch
Main Web
SpeechRetrievalTeam
English-Dutch Onlin...
IlpsSeminar
iach wiki
Kunststofzuiger – Podcast details
KUNSTstofzuiger
Motes of note from Dutch radio program Kunststof
Zoek!
(i.e. Kunst, Arie of Willem Koning)

Nausicaa Marbe en Jeroen Vullings

Uitgezonden op: **11-03-2008**

Aan de vooravond van de Boekenweek praat Kunststof een uur lang met schrijfster Nausicaa Marbe en literair criticus Jeroen Vullings (Vrij Nederland) - die in het dagelijks leven geliefden zijn - over hoe goed of hoe slecht het gaat in de wereld van de literatuur. Plus een reportage met de Amsterdamse boekhandelaar Ton Schimmelpennink. Presentatie Jellie Brouwer

Alternatieven

Sara Kroos
10-03-2008

Edwin Kats
07-03-2008

Mario Molegraaf
06-03-2008

Huub Oosterhuis
05-03-2008

nederland
(0/7)
00:23
acht actualiteit amsterdam angst antwoord bekend
boek boeken boekhandel cijfer
communicatie dag dezelfde geloof geval helft honderd
jammer jong jongeren kijk kilometer kinderen

Fig. 6.7 Player for the Kunststofzuiger Search Machine.

when the user clicks on a podcast episode that was presented in a results list in response to a query, in this case “nederland” (Eng. “Netherlands”). The player page has a query-independent representation of the episode, in the form of the podcast title, broadcast date and description, and also a term cloud that has been extracted from the transcript of the podcast. It also has a query-biased representation of the episode in the form of the player, which contains markers pointing to the moments within the podcast at which the query word occurs. These markers depict the overall frequency of the query word. Clicking one of these markers moves the user to the point in the speech stream at which the query word is spoken. Note that playback should begin a few words before the spoken word in order to allow the user to process the speech. The time lag before the spoken word should be approximately constant so that a few interactions with the system will inform the user how long to listen at a particular jump-in point before concluding that point was a false alarm and contained no mention of the keywords.

Note that the *Kunststofzuiger* search engine, whose interface is depicted in Figure 6.7, is an SCR system that takes a “finding content”-oriented approach to retrieving podcasts. However, once the content has been located, the occurrence of certain words within the podcast is made visible to the user. In other words, a“finding mentions” point of view is used to support the user in navigating within the episode.

Two variants on these strategies deserve mention. First, the term cloud can be spread out along the player to give the user a general idea of the topical development over the course of the speech media. Such an approach was proposed in [78]. Second, instead of pointers to the exact place at which a word is mentioned, a heat map can be displayed that uses shading or color to reflect the relative likelihood of a position along the timeline being relevant to a query. This approach is adopted in the VMR Broadcast News browser [26], whose browser bar is shown in Figure 6.8. In order to create this representation, the ASR transcript is divided into equal-length segments each of which is scored against the query. The brightness of the shade indicates the strength of the match of each segment, allowing identification of likely relevant regions of the broadcast. This particular example illustrates

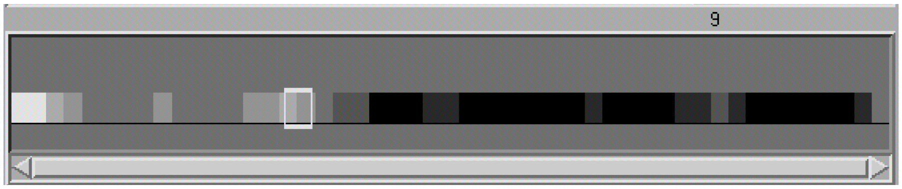
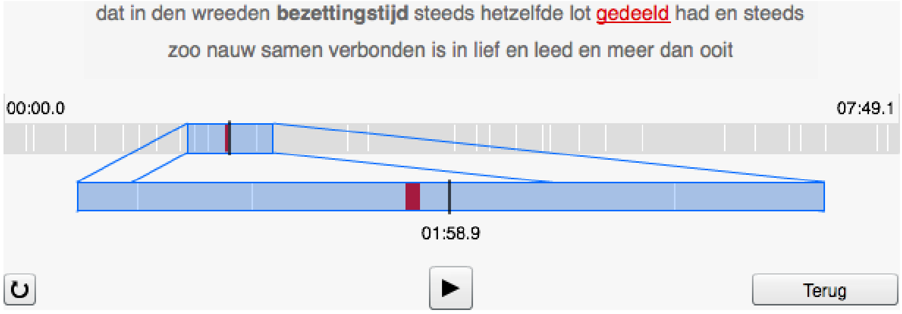


Fig. 6.8 Heat map from the VMR Broadcast News browser depicting the likelihood that regions of spoken content are relevant to the user query.



dat in den wreeden **bezettingstijd** steeds hetzelfde lot gedeeld had en steeds zoo nauw samen verbonden is in lief en leed en meer dan ooit

Fig. 6.9 Player for the Radio Oranje System.

the typical structure of a news broadcast where the story of interest is mentioned in the headlines and covered in detail in the main broadcast and then appears later when the broadcast headlines are repeated. A similar heat map approach is currently used in the commercial player called the Limelight Video Platform[1] (previously Pluggd).

In addition to depicting the position of semantic content within the speech media items, players can also represent segmentation structure. An interface that displays the whole recording together with a positional window and a simultaneous an enlargement of the positional window is described in [145]. In this case, significantly more information must be packed into the player timeline and a magnifying glass metaphor becomes useful.

In Figure 6.9, the player bar of the previously mentioned *Radio Oranje* [108, 109] application is depicted as an example of the magnifying glass metaphor. The player displays the entire speech in a timeline, as well as a magnified view showing a window of 45 seconds around

[1] http://www.delvenetworks.com

the current position of the cursor. Above the play bar, the transcript of the currently-playing segment is displayed, with the query word in bold and a moving underline tracking the progression of the playback. The magnified view makes it possible to also depict segmentation information for the entire program in a relatively compact space without losing detail.

Segmentation structure is a key characteristic of speech media used for depiction in players. Segmentation patterns act as an identifying fingerprint for speech media. A global pattern may serve to implicitly convey to the user information about the nature of the media, for example, if it is a conversational interview or a political speech. Further, displaying segments within the playback interface provides the user with the context of the results being examined, which is an important aid to interpretation. Segmentation provides an alternative for rewinding and fast-forwarding: a user can jump back to the beginning of a segment boundary. Such jumps can be considered to be "intelligent" in so far as the underlying segmentation provides a good representation of useful semantic structure of the speech media.

The most appropriate use of segmentation structure will depend on the segmentation information available, its quality and also the types of user needs and tasks the SCR system is designed to support. A large number of segments can be simultaneously displayed on the player interface by adopting the playlist metaphor and listing segments vertically [157]. Depending on the use case, space can be conserved by dropping length information and representing each segment as an equal-height line. Horizontal layout with tracks is a choice preferred for situations in which multiple overlapping segmentations exist. An early example is the PARC audio browser [76, 145], which displays separate tracks for announcer, speaker, audience, silence, and applause.

In the ideal case, manual segmentation information should be included in the metadata of the spoken content. However, this is often not the case and automatic methods such as TextTiling, presented in subsection 5.3.4, must be applied to generate segmentation boundary points. Such methods inevitably make errors, dropping real boundary markers in some places and inserting false boundary markers in others.

The utility of the browsing interface may potentially be impacted by these errors, particularly if they are numerous or occur at significant points in the semantic flow of the content.

Colors in the playback interface can be used to represent speaker characteristics such as gender and speaker age (i.e., child or adult) or even speaker identity or speech/non-speech differences [202]. The patterns of segment alternation are a potential source of valuable information that can aid the user in the selection process [157, 202, 203]. However, care must be taken, since alternating patterns may be difficult for users to grasp [253]. In general, the more immediately obvious it is to the user why the system "chose" the particular object as relevant, the more comfortable the user will feel using the interface. System designers need to pay careful attention to how the player links to the media file.

The *Radio Oranje* application in Figure 6.9, as mentioned in subsection 5.2.4, was implemented by using ASR to create a forced alignment of human-generated transcripts with spoken content. At this juncture, we present further details on several other types of alignment that can be used to improve the ability of an interface to visually represent the spoken content of a speech media item or otherwise support the user's process of reviewing or examining results. Alignment is not just limited to transcripts, but rather for any sort of speech media, if it has a parallel resource containing text, the two can potentially be aligned with the help of ASR. As previously mentioned in [130], slides of presentations are aligned with the speech media. The slides provide structure for the speech stream and act as surrogates for displaying spoken content in the interface. In [52], a system is described that transcribes broadcast news in real time, analyzes it for named entities and topic, formulates a set of queries, and uses those queries to extract information from other information sources (e.g., newspapers, WWW). Excerpts of information are inserted into the news stream in order to provide viewers with background information on the topics treated. A similar system for Dutch-language broadcast news is presented in [190]. These types of interfaces invite exploration and support the user in browsing activity, to which we now turn in more detail.

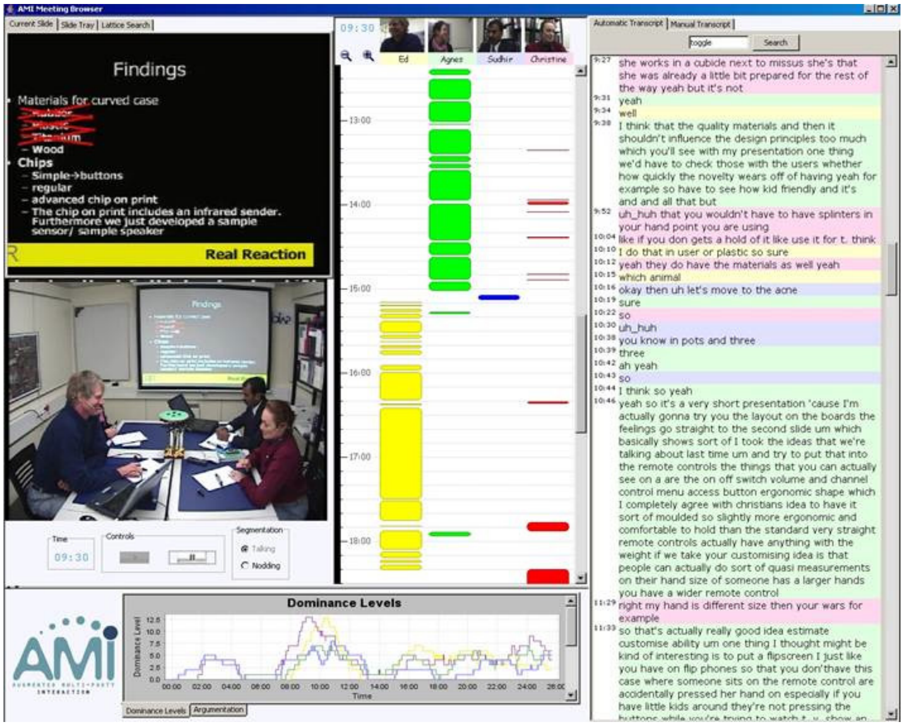
6.3.1 Browsing

The search process is only one way in which users seek information. This point was made by [203], which highlighted the important role of alternatives to the ranked list, such as spatial visualization of document collections. An example of an approach that captures the entire document collection in one structure is provided by the Chinese broadcast news prototype described in [166], which offers, in addition to a bottom up retrieval functionality, the option of top-down browsing. A list of 20 news categories is provided on the portal page. These categories provide an entry point for a 2D topic tree. Clicking on a category reveals a grid representing the latent topical structure of that category. Clicking on a grid cell reveals a finer breakdown of that category. If the user has retrieved an item via a query submitted to the system, a click on this item will reveal its position within the 2D tree structure.

Exploration is supported by a browsing interface that visualizes the structure of items and connections within the collection. Browsers that go above and beyond audio and video material, for example, by integrating slides and notes, have been designated *Artifact browsers* [299]. The design of suitable interfaces to support interaction is again crucial to the success of such systems. Note that there is not a hard boundary separating playback interfaces, discussed in the previous subsection, and browser interfaces. The difference lies in the emphasis that browser interfaces put on presenting a complete picture and on supporting discovering.

Much research effort on browsing has been devoted to the domain of meetings. Interfaces can provide access to spoken audio via time-specific links to a meeting's agenda, to images made during the meeting, for example of the whiteboard, or to automatically identify elements such as topic, functional category (presentation, discussion, break) or "hot spots" [59, 299]. We use a specific example from this domain to illustrate browser interfaces: the JFerret meeting browser developed by the AMI/AMIDA project,[2] depicted in Figure 6.10.

[2] http://www.amiproject.org/showcase/integrated-systems/meeting-archive-browsing



Findings

- Materials for curved case
  - [illegible]
  - [illegible]
  - ~~Titanium~~
  - Wood
- Chips
  - Simple→buttons
  - regular
  - advanced chip on print
  - The chip on print includes an infrared sender. Furthermore we just developed a sample sensor/ sample speaker

Real Reaction

9:27 she works in a cubicle next to missus she's that she was already a little bit prepared for the rest of the way yeah but it's not
9:31 yeah
9:34 well
9:38 I think that the quality materials and then it shouldn't influence the design principles too much which you'll see with my presentation one thing we'd have to check those with the users whether how quickly the novelty wears off of having yeah for example so have to see how kid friendly and it's and and all that but
9:52 uh_huh that you wouldn't have to have splinters in your hand point you are using
10:04 like if you don gets a hold of it like use it for t. think
10:10 I do that in user or plastic so sure
10:12 yeah they do have the materials as well yeah
10:15 which animal
10:16 okay then uh let's move to the acne
10:19 sure
10:22 so
10:30 uh_huh
10:38 you know in pots and three
10:39 three
10:42 ah yeah
10:43 so
10:44 I think so yeah
10:46 yeah so it's a very short presentation 'cause I'm actually gonna try you the layout on the boards the feelings go straight to the second slide um which basically shows sort of I took the ideas that we're talking about last time um and try to put that into the remote controls the things that you can actually see on a are the on off switch volume and channel control menu access button ergonomic shape which I completely agree with christians idea to have it sort of moulded so slightly more ergonomic and comfortable to hold than the standard very straight remote controls actually have anything with the weight if we take your customising idea is that people can actually do sort of quasi measurements on their hand size of someone has a larger hands you have a wider remote control
11:29 right my hand is different size then your wars for example
11:33 so that's actually really good idea estimate customise ability um one thing I thought might be kind of interesting is to put a flipscreen I just like you have on flip phones so that you don'thave this case where someone sits on the remote control are accidentally pressed her hand on especially if you have little kids around they're not pressing the buttons while you're trying to watch t. v. show an

Fig. 6.10 AMI meeting archive browser, JFerret.

The browser supports the user in finding elements of interest within a recorded meeting. A range of data types is displayed for this purpose, including speaker diarization, linked slides and transcripts. A curve that plots speaker dominance within the meeting is also displayed. This browser is intended as one part of a large system that would allow users to explore a corpus of meetings, for example to search for the most relevant pieces across meetings to allow the user to answer a specific question. More information on AMI/AMIDA and the JFerret meeting browser can be found in [70, 119, 292, 293].

Here, we have covered basic issues in the design of user interfaces for SCR, including query entry, results display and results playback. It can be expected that new players will make great strides over what has been shown here. Especially for mobile devices in which screen real estate is limited and in which interaction possibilities range beyond keyboard and mouse input, innovation can be expected. However, the underlying topics we have treated here, that is, graphically depicting the relevance of items and also of enabling semantic playback, can be expected to remain among the major challenges.

7

Conclusion and Outlook

The goal of this survey has been to provide an overview of research in the area of Spoken Content Retrieval (SCR). We have taken the view that SCR involves not only finding mentions of query words within the speech signal, but also relating these mentions to the user information need via a meaning-based notion of relevance. SCR can be naïvely characterized as a simple combination of Automatic Speech Recognition (ASR), which generates text transcripts from spoken audio, and Information Retrieval (IR), which identifies text documents relevant to user information needs. In this survey, we have placed special emphasis on discussing techniques and technologies that make it possible to go beyond a naïve combination and enable a tighter integration of ASR and IR. We argue that careful consideration of how ASR and IR are integrated within specific user scenarios will result in more effective SCR systems, better able to meet user needs. The survey has been organized around five key challenges of SCR that are important to address in order to achieve an optimal integration of ASR and IR for a given SCR use scenario.

After having presented a compact overview of the field of IR and introducing these five challenges in *Overview of Spoken Content*

Indexing and Retrieval (cf. subsection 2.4 “Challenges for SCR”), the survey continued, in *Automatic Speech Recognition*, with an overview of ASR technology. This overview set the background for a high-level discussion of considerations that are advisable to take into account when integrating ASR and IR (cf. subsection 3.4, “Considerations for the Combination of ASR and IR”).

The remainder of the survey presented material that delves deeper into these considerations, addressing the five key challenges of SCR. We summarize the relationship of the individual topics we have presented, and issues we have examined, to these challenges in the following summary:

- *The challenge of handling uncertainty*: In *Exploiting Automatic Speech Recognition Output*, we present techniques for using ASR output within an SCR system, covering n-best lists, lattices and confusion networks, confidence scores and also the use of subwords in SCR. All of these techniques go beyond the 1-best word-level transcript produced by the recognizer. These techniques make it possible to represent uncertainty that arises during the speech recognition process. They benefit from careful attention to considerations of normalization and pruning. Properly applied, they make it possible to deal more effectively with uncertainty during retrieval. They provide opportunities for a tighter integration between ASR and IR components, which can be optimized for particular use scenarios.
- *The challenge of covering all possible words*: Large Vocabulary Continuous Speech Recognition (LVCSR) systems make use of huge lexica containing many word entries. However, in practice it is not possible to provide the LVCSR system in advance with information concerning every word it could possibly encounter in the incoming speech signal. For this reason, the Out Of Vocabulary (OOV) problem remains a major bottleneck for ASR and also for SCR. OOV issues can be addressed by better representations of uncertainty, but we have also discussed the problem as a challenge in its

own right. Techniques that have been developed to deal with OOV include subword units, lattices and fuzzy matching. We reviewed systems that make hybrid use of combinations of these techniques and also of "searching speech" techniques such as Spoken Term Detection (STD), which can be used to extract terms from the speech signal without necessitating full LVCSR.

- *The challenge of context*: Humans produce spoken language in real-world contexts containing much more information (e.g., concerning recent events or physical surroundings) than what is encoded in the speech signal. Lack of contextual information can mean that an SCR system makes a less accurate match between user needs and speech media results. To confront this challenge, use can be made of multiple information sources that go beyond the output of the ASR system. In *Spoken Content Retrieval beyond ASR Transcripts*, we discussed how ASR output can be supplemented in order to improve SCR. In particular, we addressed the use of expansion techniques and of the exploitation of both manually and automatically generated metadata.
- *The challenge of structuring spoken content*: Many use scenarios for SCR systems involve spoken content that either lacks information about segment boundaries or is inherently unstructured in nature. In particular, lack of formal structure characterizes speech media that is produced spontaneously and recorded outside the studio or in conversational settings. Techniques that can structure spoken audio have the potential to support the recognizer in two ways: by determining the regions over which IR models should calculate relevance to user information needs and by determining the form of the results that should be presented to users. In *Spoken Content Retrieval beyond ASR Transcripts*, we presented techniques for segmentation, diarization and compact representation of speech media that can be used to address this challenge.
- *The challenge of visualization and playback*: User satisfaction with an SCR system is determined not only by the quality

of the retrieval results, but also by the way in which these results are presented. An SCR system should provide representations of spoken results that allow human users to quickly and easily ascertain which results best suit their information needs and select results to examine in more detail. In *Accessing Information in Spoken Content*, we discussed techniques for result visualization and playback that can be used to address issues of human interaction with the SCR system.

The survey has made clear that the issues faced when developing an SCR system that optimizes the integration of ASR and IR are quite substantial. SCR use scenarios can differ widely from each other (e.g., declarative vs. conversational content, planned vs. spontaneous speech, short vs. discursive queries and structured collection vs. unstructured speech streams). For different scenarios, the combination of approaches that will yield the most effective SCR system can be expected to vary. Fortunately, as reflected in this survey, the techniques and technologies available to address the challenges of SCR are numerous and quite sophisticated. The past two decades of SCR research has yielded a wealth of useful approaches, which can be drawn upon to implement SCR systems or to stimulate the development of new SCR technologies.

In the *Introduction*, four recent developments in speech media were cited: volume, variety, function, and user attitude. The importance of speech media can be expected to continue to grow, following these, or related, lines. The convergence of these developments has led to the opening of a new era of SCR. Moving forward, the driver of SCR research and development can be expected to be the user, whose information needs will determine and define new use scenarios for SCR. The goal of SCR research should be to build systems that users find genuinely useful. Motivated by their satisfaction with SCR technology, users in turn will grow more accepting of technologies that make possible spoken-content-based search in speech media.

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. Žd́ánský, 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 Commununication*, 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.