Presented at the IST Consultation Workshop on “Challenges of Future Search Engines” held in Brussels on 15th Sept. 2005.

Opportunities and Challenges in Spoken Media Search

Konstantinos Koumpis
Sira Technology Ltd.
South Hill, Chislehurst
Kent, BR7 5EH, UK
costis.koumpis@sira.co.uk

Abstract

Text search remains dominant while at the same time the generation of multimedia content is growing at a fast pace. Contrary to the past, a significant portion of this content is generated by users themselves and not media firms. Current approaches in searching and browsing speech, image and video archives are expensive and time consuming creating incentives to extend search capabilities into these media in a similar way to text. The focus of this contribution is on the field of spoken media. In particular it discusses the main application domains and identifies opportunities and challenges for the future.

1. Introduction

When someone uses a search engine to find information about a subject of interest, after entering a set of keywords, is presented with a list of potentially relevant sources, mainly in the form of web pages, documents or images. Yet until recently, search engines were unable to process information preserved in the form of speech – interviews, lectures, voicemails and radio newscasts – even though this medium allows easy and affordable storage of millions of hours of human knowledge and interactions. Spoken media content combines information from multiple levels (phonetic, acoustic, syntactic, semantic and discourse) and is far richer than what a simple textual transcription can capture. For instance, it has additional cues that disclose the intended meaning and speaker’s emotional state.

Spoken media search has become feasible thanks to a number of advances fueled by scalable statistical models with efficient algorithms for inference and decoding, increases in computational resources and the development of large, annotated databases. Systems for content-based spoken media search are typically built using separately developed spoken language processing and information retrieval (IR) components (cf. [1] for an overview). The speech recognition component converts the input speech signal into word sequences. It must handle continuous speech and be speaker independent, eliminating the need for it to be pre-trained for individual speakers. Real-time operation and high accuracy are not as strict requirements as is the ability to handle large amounts of pre-recorded or streaming data. Since speech recognition systems can label automatic transcriptions with exact time stamps, their output can be viewed as a form of annotation with which the other tasks can synchronise (Figure 1). Topic segmentation/tracking and speaker detection/tracking are used as a basis for indexing relevant audio segments according to topic or speakers, respectively. Specific information, such as named entities (NE), can also be extracted automatically from the transcriptions to facilitate more detailed analysis.

In the retrieval phase, the focus is on selecting which terms from the text and annotations to compare, how they should be weighted, and how to compare the sets of weighted terms. An advantage of re-usable annotations is that they can be used to infer additional semantic relations. Retrieval can also be facilitated by classifying content into categories. This not only simplifies the retrieval process itself but can also assist users to better understand and remember information as it is presented in the appropriate context. The last and perhaps the least explored phase deals with the delivery of the retrieved content to users. Summarisation is a promising method to overcome the problems associated with information overload by presenting condensed versions of the content. User interfaces typically supports queries expressed in natural language or with Boolean expressions. Adaptive profiles that tend to reflect long term information needs can also be used to replace repeated queries and filter our irrelevant information.

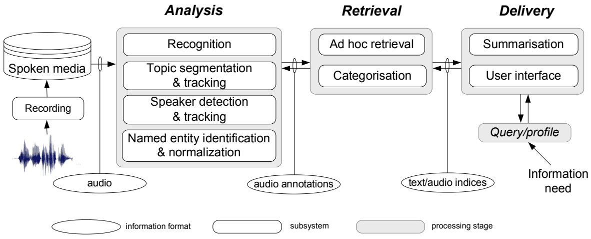

Figure 1. A generic system architecture for spoken media search and browsing.

The rest of this contribution is organised as follows. In section 2 the application domains that have attracted significant research interest are briefly reviewed. Section 3 presents the main opportunities created by the recent technical advances and the present market dynamics, while section 4 lists eight research challenges. The contribution is concluded in section 5.

2. Current Application Domains

The application domains that have attracted the most interest in the field of spoken media search are newscasts, voicemail and conversational speech. A major difference across these domains is the quality of automatically generated transcriptions, which varies from

10-20% word error rate (WER) in newscasts to 20-40% WER in voicemail and conversational speech.

Newscasts

The newscasts domain has attracted a lot of interest since it is very general, allows relatively easy data collection and the market demand is well defined (i.e. content management needs of large newscast firms). This domain involves a wide variety of speaking styles (e.g. reporters, politicians, common people and news anchors) over high-quality microphones but also some interview reports which are transmitted over a telephone channel with a reduced bandwidth and background noise, or overlapping speakers. Recognition of proper names and unknown words is problematic in this domain and as such phone-based or keyword spotting approaches have been considered. Tuning the vocabularies to specific collections and time periods requires additional effort and automatic techniques have been proposed. A number of retrieval systems, operating on archives of newscasts, were evaluated as part of the Text REtrieval Conference (TREC), giving the important result that retrieval performance on automatic speech recognition output was similar to that obtained using human-generated reference transcripts, with little or no dependence on transcription errors. Comparable results have since been achieved across several languages. It has also been found that the accuracy of the NE identification task (about 10% of the transcribed words in newscasts are NEs) is strongly correlated with the number of transcription errors.

Voicemail

The domain of voicemail involves a conversational interaction between a human and a machine with no feedback from the machine. Voicemail messages are typically short, conveying the reason for the call, the information that the caller requires from the voicemail recipient and a return telephone number. Manual organisation of voicemail is a time consuming task, particularly for high-volume users. A few alternative solutions have been proposed for efficient voicemail retrieval. The ScanMail system supports browsing of message transcriptions via a graphical user interface. The VoiSum system proposed the generation and delivery of text summaries on mobile phone displays by extracting content words from the message transcriptions using combinations of lexical and prosodic features.

Conversational Speech

Conversational speech in unrestricted domains is very challenging due to its spontaneous nature and the need for multi-speaker processing (speaker activity and overlap detection). Speech in such domains tends to be more complex than that used in human-to-machine interactions, showing complex syntax, more words per utterance, and more ambiguity. The DiaSumm system has addressed some dialogue-specific issues of summarisation such as disfluency detection and removal, sentence boundary detection and topic segmentation. Efforts are also under way to analyse large multilingual interviews containing spontaneous, accented, emotional and elderly speech as part of the MALACH project. Apart from the technical obstacles, a number of socio-cultural issues, such as privacy are of higher importance in conversational speech rather than the other domains.

3. Opportunities

Only a few years ago, search did not seem to matter that much. Today though, its impact on information accessibility and integration is profound. As a result, search has evolved from an enabling technology to a business enabler and market differentiator. The increasing volume of spoken media including material generated by end-users creates strong economic incentives to port components that support spoken media search to personal archives. The search business is still young, and could fragment into many different products for different purposes. However, early comers with an established base of registered users are likely to have a competitive advantage in the future. Some examples of potentially important applications that span from individual to government levels are listed in the table below.

Table 1 Examples of key applications of spoken media search.

Individual	Personalised delivery of voicemail and newscasts
	Search in audio books and music (incl. similarity search)
	Search in personal audio recordings (meetings, presentations, telephone conversations)
Education	Access to audio visual material for training and research purposes
Corporate	Retrieval of help desk calls
	Content management of corporate meetings and conference calls
	Monitoring of television and radio news
Government	Access to audio proceedings (parliamentary sessions, court of law archives)
	Access to cultural heritage archives
	Monitoring of unlawful conversations for security purposes

The main drivers and early adopters of spoken media search technologies are expected to be the providers of telecommunication services and the major Web search engines[1], followed by the more traditional media companies such as music, television and film content providers. The explosion of user-generated media content creates significant opportunities for providers of consumer electronics too.

Telecoms Providers

Traditional telecoms providers, up to 90% of whose revenues come from fixed-line calls, are already looking for ways to add value to their services and compete with the free but basic peer-to-peer calling based on the Internet protocol (IP). Telecoms providers will need to offer new services that are enabled by content-based access to spoken media such as voicemail management, real-time language translation and indexing of phone conversations. For these services privacy issues will be a critical barrier to adoption.

[1] The difference between these type of service providers may become blurred in the future given that several Web search engines have started providing Voice over IP telephony services.

Web Search Engine Providers

The Web portal model that provides users with links to whatever pages the portals' editors consider useful and generates revenue by presenting advertisements and encouraging users to sign up for additional services, has been significantly challenged by a new breed of powerful search engines. As of today though, these web search engines have not supported spoken media search, mainly because of the technical difficulties in it in comparison to hypertext files. Nevertheless, web search engine providers are eager to extend their offerings from the domain of hypertext to multimedia content in general. As an intermediate step, it is possible to provide basic spoken media search services without performing content analysis (for some types of content at least), for instance by using the information found in simple annotations that are already available (e.g., file headers), such as producer, length or date. Yet another possibility would be to exploit the associations between spoken media files.

Media and Entertainment Providers

Media providers have increasing needs for efficient management of the materials they produce. Spoken media search firstly allows their content editors to access related stories quickly and secondly gives them the capability to offer subscription-based access to their archives to interested individuals and corporations.

Consumer Electronics Providers

The proliferation of personal devices for capturing and reproducing sound and images (e.g. personal digital assistants, digital cameras, mobile phones, music players, video recorders) has resulted into a rapid increase of user-generated media. Users may wish to search their own archives or those of their peers for similar audio files according to a number of attributes (e.g. topic, speaker, date, popularity, and types of background noise). A new breed of approaches may be required for this type of search as the ranking techniques that have proven very effective in hypertext search are not applicable to personal archives (e.g. podcasts, photo and video blogs) where most content is not hyperlinked.

4. Research Challenges

In order for applications such as those listed in Table 1 to be successfully realised, research is needed in a number of areas. Although the list below is not meant to be exhaustive, it covers some of the most problematic technical areas. A number of legal (e.g. copyright) and social (e.g. privacy) issues are very crucial, but beyond the scope of this contribution (cf. [2] instead).

Annotation and Semantics

Annotation tags can be used as a basis to perform semantic search so as to extend and improve traditional search processes based on IR technology. However, most multimedia content does not have a lot of useful tags around it. Companies or individuals wishing to make their content searchable today have to go through the expense and effort of detailed extensible markup language (XML) tagging and resource description framework (RDF) building. This procedure needs to be simplified possibly by incorporating ontologies – structured sets of concepts with agreed relationships that represent real-world knowledge. Ontologies will enable search for meaning rather than words and make the content more accessible. This is particularly useful given the fact that people use many synonyms and express things like dates and locations in many different ways. Ontologies that are derived automatically from annotations will enable users to perform a single semantic search to retrieve all the relevant information about a topic.

Unified Modelling Approach

Despite the diverse role of subsystems for content analysis, retrieval and delivery, the majority of them are approached with the same perspectives and modeled using the same or similar statistical frameworks, namely Markovian. However, this fact has not yet been translated into a unified modeling approach, and as a result the trainability and scalability of the component models remains limited. Furthermore, most applications are driven by application-dependent heuristics. If this trend continues, there is a risk of failing to support very large spoken media archives or keep making advances in tasks more demanding than retrieval. More compact system architectures resulting from a unified modeling approach would also play a major role in model validation and portability to new domains.

Non-verbal Cues

Spoken media search and browsing tasks can be significantly benefited from a systematic integration of prosodic cues, which are largely ignored despite being essential components in the way humans structure their intent and mediate interpretations in context. For instance in an extractive voicemail summarisation task we found that combined lexical and prosodic features were up to 10% more robust than combined lexical features alone. At the same time, integration of cues from video processing (e.g. gestures, speaker localisation) in selected domains where audio-visual data can be obtained will reduce the ambiguities during audio content analysis, as caused by background noise, poor recording or overlapping speakers.

Categorisation

Users often do not use correct keywords in their queries. The goal of automatic categorisation is to assign segments of spoken media or their transcriptions to relevant categories. Manual construction and maintenance of rules for categorisation is a labour intensive and possibly unreliable operation. It is possible instead to build classifiers automatically by learning the characteristics of the categories from a training set of preclassified examples. Many standard machine learning techniques have been applied to automated text categorisation problems, such as decision trees, naive Bayes classifiers, k-nearest neighbor classifiers, neural networks and support vector machines. However, only a limited amount of work on automatically generated transcriptions has taken place.

Mobile Access

As users increasingly prefer to access content using wireless handheld devices, the associated application design implications of mobile access should be considered for spoken media search and browsing too. Data entry using a keypad should be kept to a minimum given that users may need to access content while they are walking or driving. In applications where simple but fast task completion (e.g. retrieval of newscasts) is required, user profiles that adapt over time and tend to reflect long-term information needs can be employed instead of repeated queries. Profiles allow content access in context (what have you seen/heard, where you have been). Advances in the analysis and retrieval tasks will allow user interfaces to support natural text or speech input (e.g. questions), or support for providing samples of spoken media examples (e.g. related to a speaker or background conditions). Mobility may also impose specific challenges on search and browsing due to the diversity of networks and platforms.

Summarisation

Humans can assimilate information faster through the eyes than the ears and empirical studies have suggested that summaries can save time in digesting audio content. Speech summarisation reduces the size of automatically generated transcripts in a way that retains the important information and removes redundant information. Although there has been much research in the area of summarising written language, a limited amount of research has addressed the creation and evaluation of spoken language summaries based on automatic transcriptions. A complete speech abstraction system, that generates coherent summaries by paraphrasing content, demands both spoken language understanding and language generation and is beyond the current state of the art. However, it is possible to use simpler techniques to produce useful summaries based on term extraction, sentence extraction/compaction and concatenation. This task is based on selection of original pieces from the source transcription and their concatenation to yield a shorter text. A major advantage of the extractive summarisation approach in comparison to abstraction is its suitability for supervised training and objective evaluation given the existence of example summaries.

User Interfaces

A good user interface is easy to use, attractive to the user and offers instant feedback. The choice between content delivery via text or via audio should take into account the characteristics of the content, such as its duration and operating environment as well as the limitations of human cognitive processing. Early spoken media content access systems such as Scanmail, SpeechBot, Rough'n'Ready and THISL followed the dominant paradigm established by Web search engines with which both the designers and the potential users were familiar. Queries were primarily expressed as typed text, while the output was enhanced text displayed on a screen. Because the automatically generated transcripts contain recognition errors, to support a final decision systems typically provide users with the ability to playback segments of individual recordings. This paradigm became known as "what you see is almost what you hear" emphasising the inevitability of transcription errors. Such "less than perfect" searches are acceptable in many situations though and user interfaces can be cleverly used to hide technology imperfections. Over the last few years, user interfaces for accessing spoken media content have started addressing on a number of other high-level user interface issues, such as topic segments and speaker turns, construction of audio scenes and presentation of non-verbal information.

Revenue Model

Cool demos do not lead to adoption and commercial success. Finding workable revenue models and demonstrating return on investment will be major challenges for both providers and adopters in the field of spoken media search. This is due to the high development effort and unproven business value – for some categories of adopters at least. Advertising has helped pay for the Web-based free search services. But given the cost of recording speech and the complexity of analysing it, revenues even from contextually targeted advertising might not be satisfactory. The industry will have to design and prove in practice how the various pay per use and subscription based revenue models will be of value to users willing to pay for their searches. Ultimately, success will be determined by the ability to deliver content that the end-users want and that content providers will consider economically viable to make available.

5. Concluding Remarks

Search engines add structure to content after it has been created and this structure results into added value as a result of increased accessibility. The underlying technologies thrive in the conditions that support the information revolution – high levels of affordable storage capacity, computing power and universal connectivity. A major new frontier for search engines is to provide effective access to spoken media content in a similar way to text. As the field of text search is maturing, major advances are inevitable in the field of multimedia search as a result of resource reallocation and data availability. The diversity and complexity characterising the main research challenges reviewed above as well as other issues such as privacy and copyright suggest that more years of R&D will be required before highly effective systems become available. In the mean time, a system does not have to be perfect to be useful. Success will be ultimately determined by the ability to deliver content that the end-users want and that content providers will consider economically viable to make available.

References

[1] K. Koumpis and S. Renals. Content-based Access to Spoken Audio, *IEEE Signal Processing Magazine (Special Issue on Speech Technology and Systems in Human-Machine Communication), Vol. 22, No. 5, pp. 61-69, Sept. 2005.*

[2] J. Goldman, S. Renals, S. Bird, F. de Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. Oard, C. Stewart, and R. Wright. Transforming Access to the Spoken Word. *International Journal of Digital Libraries*, 2005. In press.