Content-based access to spoken audio

Content-based Access to Spoken Audio

Konstantinos Koumpis and Steve Renals

The amount of archived audio material in digital form is increasing rapidly, as advantage is taken of the growth in available storage and processing power. Computational resources are becoming less of a bottleneck to digitally record and archive vast amounts of spoken material, both television and radio broadcasts and individual conversations. However, listening to this ever-growing amount of spoken audio sequentially is too slow, and the bottleneck will become the development of effective ways to access content in these voluminous archives. The provision of accurate and efficient computer-mediated content access is a challenging task, because spoken audio combines information from multiple levels (phonetic, acoustic, syntactic, semantic and discourse). Most systems that assist humans in accessing spoken audio content have approached the problem by performing automatic speech recognition, followed by text-based information access. These systems have addressed diverse tasks including indexing and retrieving voicemail messages, searching for broadcast news, and extracting information from recordings of meetings and lectures. Spoken audio content is far richer than what a simple textual transcription can capture as it has additional cues that disclose the intended meaning and speaker's emotional state. However, the text transcription alone still provides a great deal of useful information in applications.

This article describes approaches to content-based access to spoken audio with a qualitative and tutorial emphasis. We describe how the analysis, retrieval and delivery phases contribute making spoken audio content more accessible, and we outline a number of outstanding research issues. We also discuss the main application domains and try to identify important issues for future developments. The structure of the article is based on general system architecture for content-based

access which is depicted in Figure 1. Although the tasks within each processing stage may appear unconnected, the interdependencies and the sequence with which they take place vary.

Since speech recognition systems can label automatic transcriptions with exact time stamps, their output can be viewed as an annotation with which the other tasks can synchronize. 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 be extracted automatically from the transcriptions.

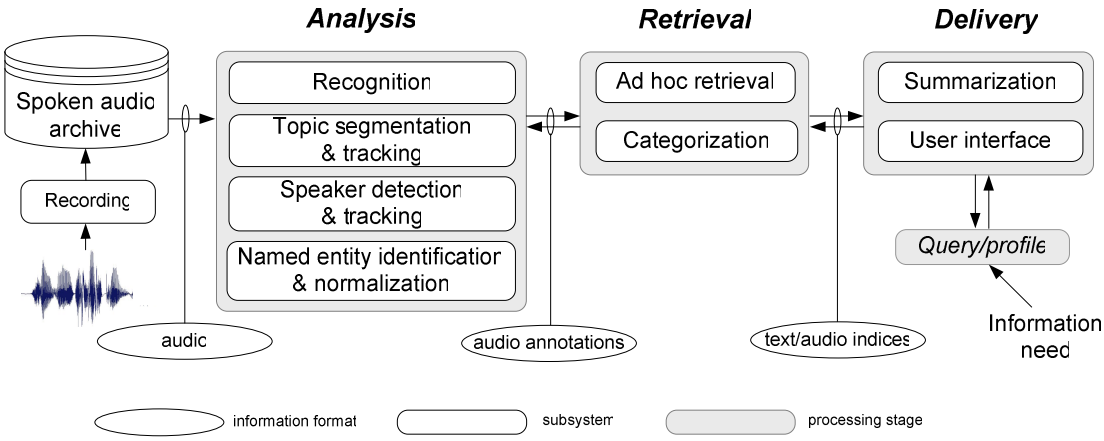

Figure 1. Generic system architecture for content-based access to spoken audio.

In the next phase of content retrieval the focus is on selecting which terms from the text and metadata to compare, how they should be weighted, and how to compare the sets of weighted terms. One way to facilitate retrieval is by classifying content into categories. The last and perhaps the least explored phase deals with the delivery of the retrieved content to users. Summarization is a promising method to overcome the problems associated with information overload by presenting condensed versions of the content. The interface 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.

Content Analysis

Analyzing spoken audio is a prerequisite for making its content accessible. Spoken language is characterized by disfluent phenomena such as incomplete sentences, hesitations and repetitions, all of which can complicate its analysis [28]. Natural speech can also change with differences in global or local speaking rates, pronunciations of words within and across speakers and different contexts. Other factors that affect the speech signals include room acoustics, channel and microphone characteristics and background noise. Although humans have developed mechanisms to compensate for the above phenomena, most of them are still very challenging for machines. These factors make the analysis of audio content a topic of ongoing research.

Recognition

Speech recognition, the task of converting the input speech signal into word sequences, is most often associated with systems for command and control, or for dialogs in limited domains. However in content-based spoken audio analysis it is employed in the form of a large vocabulary continuous speech recognizer (LVCSR) [9]. A fundamental difference between LVCSR and speech recognizers used in dictation and command-and-control tasks (like speech interfaces to web browsers and telephone banking) is that real-time operation and high accuracy are not as crucial as the ability to handle massive amounts of pre-recorded or streamed audio data.

Today's most effective speech recognition approaches are based on statistical models of small units of speech, as depicted in Figure 2. A conventional approach to front-end signal processing (see [19] for an alternative) results in a feature vector

derived from the power spectral envelope computed over a short window (20-30ms), with a typical interframe step size of 10ms. The speech signal is matched against an acoustic model, which encodes the acoustic realization of the speech. This acoustic model typically has the form of a set of stochastic finite state machines, or hidden Markov models (HMMs) [21]. HMMs for speech recognition comprise an interconnected group of states that are assumed to emit a new feature vector for each frame according to an output probability density function associated with each state. The topology of the HMM and the associated transition probabilities provide temporal constraints. Following a number of simplifying assumptions, state sequences within a HMM can yield acoustic sequence likelihoods. Speech recognition proceeds by combining these likelihoods with prior probabilities for word sequences (the language model) leading to a choice of the word sequence hypothesis with the maximum posterior probability given the models and the observed acoustic data. LVCSR systems commonly use an *n*-gram language model ([*n*-1]th order Markov model), where *n* is typically 4 or less [24]. A finite vocabulary defines the set of words or phrases that can be recognized by the speech recognizer. The size of the recognition vocabulary plays a key role in determining the accuracy of a system by introducing a trade-off between coverage and robustness of model estimates.

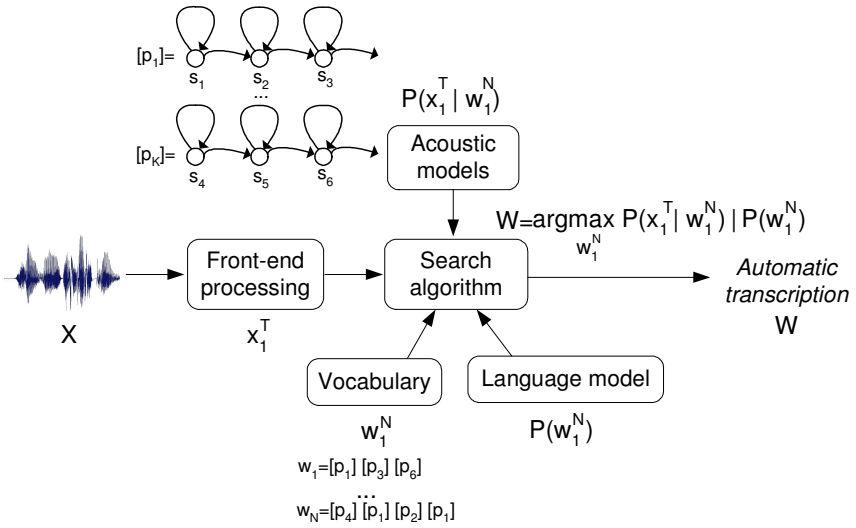

Figure 2. The dominant paradigm in automatic speech recognition, using statistical models which are trainable and scalable.

Recognition performance is typically measured in terms of the word error rate (WER), defined as the sum of the insertion, deletion and substitution errors between the recognized and desired word strings, divided by the total number of words in the reference string. Recognition performance is highly dependent upon the availability of sufficient training materials for the languages and audio data types of interest. State-of-the-art LVCSR systems are typically trained with several tens to thousands of hours of audio and several hundred million words of text and their WER varies significantly across domains. Although planned, low-noise speech (such as dictation, or a news bulletin read from a script) can be recognized with a WER of less than 10%, conversational speech in a noisy or otherwise cluttered acoustic environment or from a different domain may suffer a WER in excess of 40%. Speech recognition can be improved by speaker/condition adaptation and efficient algorithms exist for this. If real-time recognition is not a strict requirement, confidence-tagged alternative word hypotheses can be compared or hypotheses generated from various recognizers can be combined to reduce the WER.

Topic Segmentation and Tracking

A topic refers to an event or activity of interest and topic segmentation provides a high level structuring of content according to the topics covered in its segments (Figure 3, top). Performance is measured as a linear combination of the system's missed detection rate and false alarm rate which is typically presented in detection error trade-off (DET) plots. The task of topic segmentation has attracted much attention as part of recent evaluations on spoken news data, but remains far from solved for speech in unrestricted domains.

As a first processing step, speech activity areas might be automatically identified and non-speech segments (noise or music) removed. Most approaches to topic segmentation have been based on statistical models at the word level and have treated the problem as one of modeling topic boundaries, typically using maximum entropy [1], or of modeling coherent segments of text [11]. The boundary modeling approach has been successfully applied to speech [6], and this framework is suitable to extend the model to include prosodic features which are observed in the energy, intonation and timing of speech [27]. Although a number of studies have revealed that pause duration is a good predictor for topic boundaries, more experiments are needed to understand its role in spontaneous speech. This is because speakers may pause while changing their mind about what they want to say, or fill a pause while they are planning their next utterance. Approaches to topic segmentation, which frame the problem as one of classification (albeit with unbalanced classes), are quite general and have been applied successfully to similar segmentation problems including sentence boundary detection and automatic capitalization.

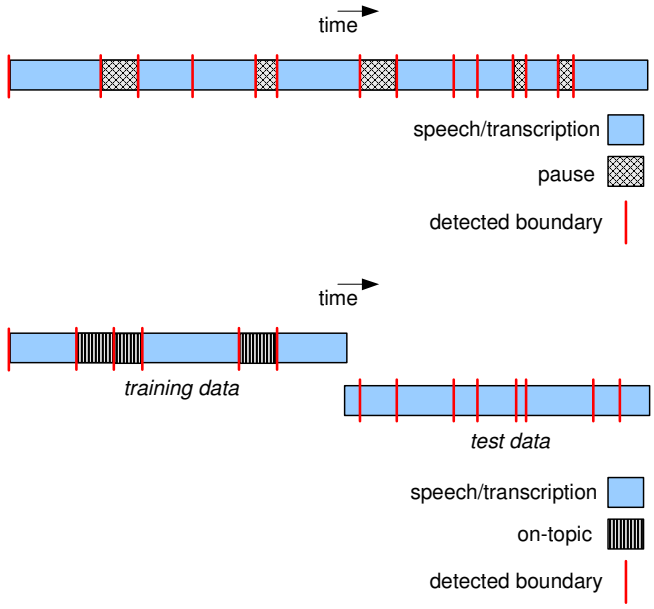

Figure 3. Topic segmentation (top) and tracking (bottom) in spoken audio.

After the topics have been segmented and identified the task of topic tracking deals with how these topics develop over time. This task is approached using supervised training given a number of sample stories that discuss a given target topic. The goal is to find all subsequent stories that discuss the target story (Figure 3, bottom). The core of most approaches to topic detection is computing term overlap between different segments: the more common terms, the more likely those two segments have the same topic. As with other text-based analysis tasks either vector space approaches or statistical language models can be used. A variation of the topic tracking task known as adaptive filtering involves detection of stories that discuss the target topic when a human provides feedback to the system (on or off-topic).

Speaker Detection and Tracking

In multispeaker audio, the association of speech segments with individual speakers is of great importance, for instance, in annotating meeting recordings or retrieving the segments in spoken news associated with individual speakers. It can also be used

prior to applying speaker adaptation techniques, as these require segments from only one speaker.

Speaker detection is possible in theory since each utterance from an individual is produced by the same vocal tract, with a typical pitch range and a characteristic articulation. However, in practice this is a very hard task as the characteristics of a given individual's voice change over time and depend on his or her emotional and physical state as well as the environmental conditions. Approaches based solely on the transcribed content of speakers' utterances have been investigated, since speakers use characteristic vocabularies and patterns of expression. Determining a speaker's identity based on transcriptions alone is nevertheless a far more difficult task for both people and machines [7].

Traditional approaches to speaker recognition are designed to identify or verify the speaker of a speech sample. The problem is treated in a similar way to that of speech recognition, typically employing typically Gaussian mixture models and HMMs [5]. For content access purposes, the basic recognition approach needs to be expanded to handle both detection and tracking of speakers in multispeaker audio.

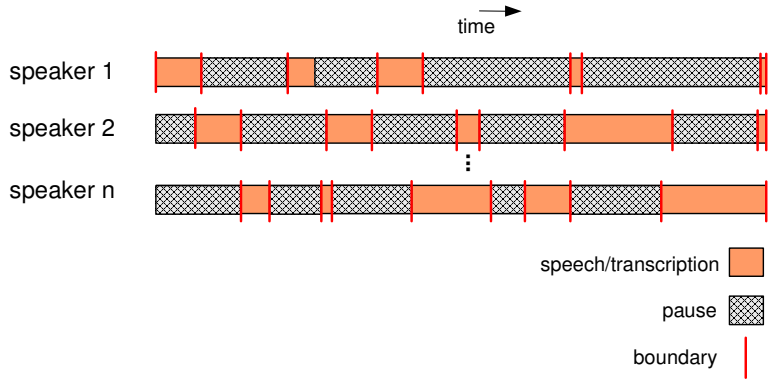

Figure 4. Speaker detection and tracking in spoken audio.

Given an audio file containing multi-party conversations and a hypothesized speaker the task of detection is to determine if the hypothesized speaker is active in the audio file (Figure 4). The speaker is detected by comparing the input speech with the

speaker models constructed in advance. Performance is evaluated, as for the topic segmentation and tracking task, in terms of the detection errors, misses and false alarms. When the audio file contains speech from one speaker a likelihood ratio statistic between a model of the hypothesized speaker and a background model representing the alternative hypothesis is computed using all available speech. In the case of more than one speaker, the speech stream is segmented into homogeneous segments (often by assuming that speakers are not active simultaneously) and then obtain likelihood ratio scores are computed over these single-speaker segments. Apart from reducing the manual effort required to track speakers throughout individual recordings, speaker tracking can potentially allow previously unknown speakers to be located in a large audio archive using a sample of speech.

Named Entity Identification and Normalization

The task of identifying named entities (NEs) is to identify those words or word sequences that denote proper names, places, dates and times and certain other classes such as numerical values. NEs are most common in spoken news, where they account for about 10% of words. NE identification is not a straightforward problem. While *Monday the Twelfth of August* is clearly a date, and *Alan Turing* is a proper name, other strings, such as *the day after tomorrow* and *Nobel Prize* are more ambiguous.

Both rule-based [31] and statistical [3] approaches have been used to perform NE identification. Rule-based approaches use grammars, gazetteers of personal and company names, and higher level aids such as the identifying co-referring names. Purely trainable systems with NE-annotated corpora can be based on ergodic HMMs (in which each state is reachable from other state) where the hidden states corresponded to NE classes and the observed symbols correspond to words. Such systems are similar to HMM-based part-of-speech taggers. A single NE can be

uttered and transcribed in several different forms. NE normalization is used to solve this problem by removing variation and mapping NEs to a single consistent root. As the co-references are found, and the different forms of a name are unified, relationships between entities can also be defined in an attempt to improve retrieval through more compact indexing.

Content Retrieval

Following the phase of analyzing spoken audio into transcriptions and metadata its content can be retrieved. This can be performed with either *ad hoc* retrieval or categorization.

Ad hoc retrieval

The task of *ad hoc* retrieval is to return a set of those audio segments or their transcriptions, judged to be relevant to a query. This task is commonly treated using approaches borrowed from text retrieval along with performance measures such as recall and precision. Reasonable results can be achieved by using simple term weighting approaches such as the term frequency inverse document frequency (tf.idf) scheme [25] across segments and the overall corpus. Frequent non-content words (e.g., 'a', 'the', 'to') are typically excluded from the retrieval models because they add little value when searching. Suffix stripping and subsequent mapping to a common root typically improves the retrieval results. For instance, the words *compute*, *computer*, and *computing* can easily confuse a speech recognizer, but given that their semantic function is similar, these words can be mapped to the single stem '*comput*'. Another method is based on query expansion, searching with additional orthographic variants and semantically related terms (perhaps derived from a thesaurus). Query expansion may also use acoustic similarity in the form of phone

lattices to account for possible errors in the speech recognition phase. Another way to improve retrieval results is to incorporate relevance feedback, assuming that the user is probably in good position to judge the relevance (or irrelevance) of a returned segment. Requerying after adding all terms from 'relevant' segments and removing of all terms from irrelevant segments is generally an effective approach to increase retrieval precision.

Statistical language models have also been applied to spoken document retrieval [20]. With this approach each segment's transcription is viewed as a language sample and the probabilities of producing the individual terms in a segment are estimated. A query is then assumed to be generated by the same process. Given a sequence of terms in a query, the probabilities of generating these terms according to each segment model are computed. Combining these yields a ranking of the retrieved segments: the higher the generation probability, the more relevant the corresponding segments to the given query.

Categorization

Users often do not use correct keywords in queries. The goal of automatic categorization is to assign segments of spoken audio or their transcriptions to relevant categories. This not only simplifies the retrieval process but can also assist users to better understand and remember information as it is presented in the appropriate context. Manual construction and maintenance of rules for categorization is a labor 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 pre-classified examples.

Many standard machine learning techniques have been applied to automated text categorization problems, such as decision trees, naive Bayes classifiers, k-nearest

neighbor classifiers, neural networks and support vector machines [26]. These approaches are effective when the segments to be categorized contain sufficient numbers of category specific terms allowing term histogram vectors (so-called 'bags-of-words') to distinguish among the categories. In this method the number of occurrences of each term matters, but the order of the terms is ignored. Stochastic language models can partially overcome this limitation by incorporating local dependencies and thus preserving more stylistic and semantic information by modeling term sequences.

Content Delivery

Spoken audio content may be delivered in either auditory or textual form. Depending on its nature (e.g. length, number of speakers) and the intended uses of the content, users may prefer to listen to a segment of the original audio recording and/or read its transcription. The differences in human capabilities in processing speech versus text will often determine the most appropriate form for content delivery. The auditory form preserves the acoustic information present in the original audio segment, disclosing the intended meaning and speaker's emotional cues, but its sequential nature makes it hard to extract information. On the other hand, speech content in textual form can be easily displayed on screens for reading, but inevitably contains transcription errors and lacks the nonverbal information that could help in disambiguating the meaning. The following section describes speech summarization, an emerging field that addresses some of the limitations of having spoken audio content in textual form.

Summarization

Speech summarization reduces the size of automatically generated transcripts in a way that retains the important information and removes redundant information

analogous to summarization of text documents. Although there has been much research in the area of summarizing written language, a limited amount of research has addressed creating and evaluating 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 (depicted in Figure 5) 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 summarization approach in comparison to abstraction is its suitability for supervised training and objective evaluation given the existence of example summaries.

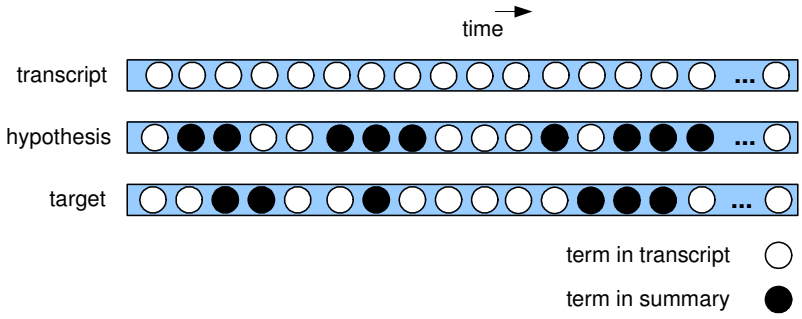

Figure 5. Extractive summarization of spoken audio.

Two distinct types of summarization tasks have been studied: (a) condensing content to reduce the size of a transcription according to a target compression ratio, and (b) presenting spoken audio retrieval results. Depending on the nature of the content and the user information needs, the processing units for summarization are either single content words or longer phrases. The features used to identify the most relevant segments from the transcription have been linguistic significance (the likelihood that the extract carries important information) [13, 15], acoustic confidence

(the likelihood that the extract has been transcribed correctly) [13, 15], and prosody [15, 33]. Methods that have been used to select the most relevant segments include maximal marginal relevance (MMR), lexical chaining, maximum realizable receiver operating characteristic (MRROC) and finite state transducers.

There are plenty of parameters to consider in evaluating summaries as various kinds of comparisons can be involved (e.g., system summaries compared with human summaries, full-transcription or system summaries compared with each other), but empirical studies have suggested that summaries can save time in digesting audio content. Note that the use of speech summarization does not necessarily imply delivery of content in textual form. It is possible to convert the text summary back into a speech signal suitable for listening using a speech synthesis or voice conversion system [30] or by processing and concatenating the relevant segments of the original audio.

User Interface

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. A good user interface is easy to use, attractive to the user and offers instant feedback. Early spoken audio content access systems such as Scanmail [12], SpeechBot [29], Rough'n'Ready [17] and THISL [22] 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" emphasizing the inevitability of transcription errors.

Over the last few years though, user interfaces for accessing spoken audio content have been designed in a way to concentrate on more complex phenomena. Figure 6 illustrates an exemplary purpose-designed user interface [32], which allows users to browse meeting recordings and quickly retrieve and playback segments of interest. Various kinds of data are displayed simultaneously, along with the video, audio, slides and whiteboard content. A user can choose a meeting to watch and which data types to display. Apart from speech transcriptions, both topic segments and speaker turns are available.

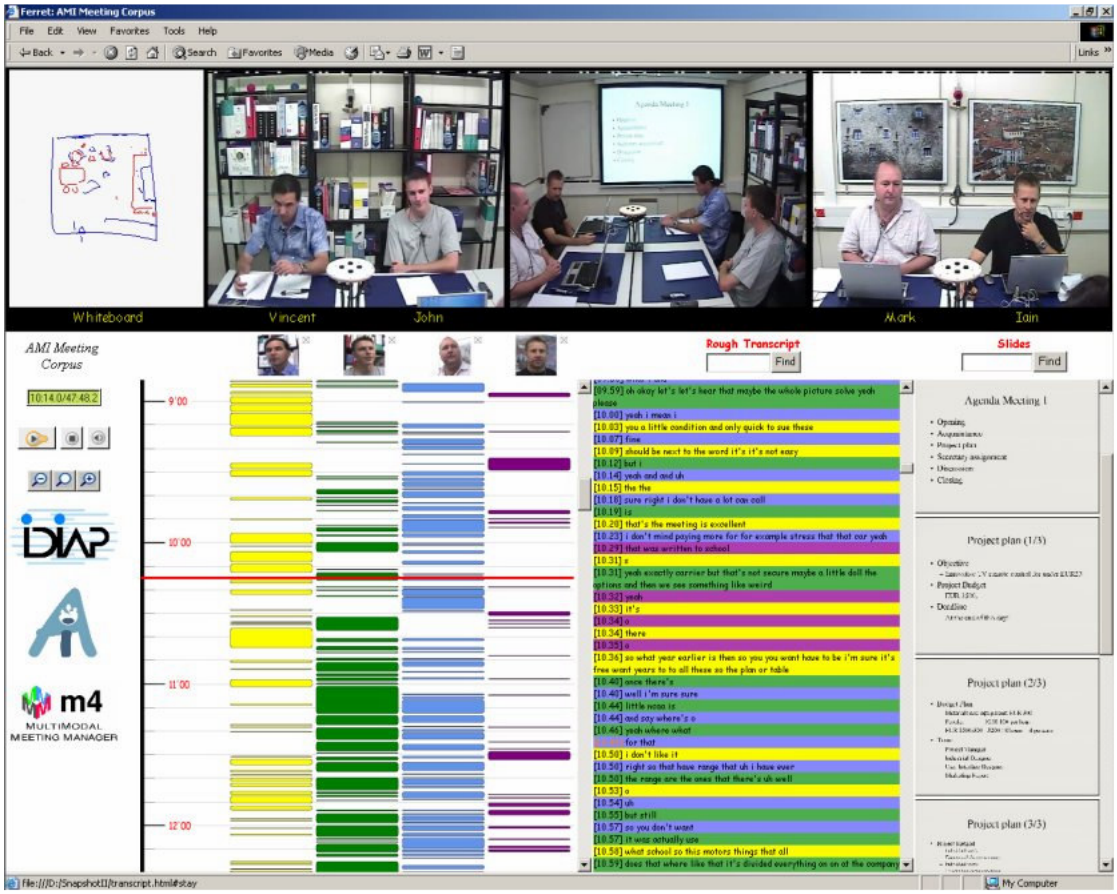

Figure 6. A snapshot of the user interface of the Ferret browser [32] *which allows interactive browsing of meeting recordings. (Figure used with permission).*

Since the most suitable is largely task-dependent (whether producing and archiving spoken news or analyzing meeting recordings), the evaluation of the overall human-system performance is critical to its selection, given all the constraints. In this

respect, the design of a good interface must take into account the system performance, the user requirements and the particular task in order to find a good match between them. There is a growing knowledge of best practice for user interface design while a number of other high-level user interface issues, such as personalization, construction of audio scenes and presentation of nonverbal information in speech are attracting research interest.

Application Domains

The most prominent application domains where techniques for accessing content in spoken audio have been applied are spoken news, voicemail and conversational speech. A major difference among these domains is the quality of automatically generated transcriptions, which varies from 10-20% WER in spoken news to 20-40% WER in voicemail and conversational speech.

The spoken news domain involves a wide variety of speaking styles (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 often include background noise, or overlapping speakers. This domain has attracted a lot of interest since it is very general, allows relatively easy data collection and offers a clear path to commercialization. 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 [2]. A number of retrieval systems, operating on archives of spoken news, 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

[8]. Comparable results have since been achieved across several languages other than English. Although, WER of up to 40% can be tolerated in terms of retrieval performance, usability studies have shown that the transcription errors affect the overall performance as perceived by users. It has also been found that the accuracy of the NE identification task (about 10% of the transcribed words in spoken news are NEs) is strongly correlated with the number of transcription errors [16].

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 organization 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 [12] supports browsing of message transcriptions via a graphical user interface. Hand-crafted rules, grammatical inference of transducers and classifiers using a set of *n*-gram features were compared within the task of extracting of the identity and phone number of the caller from voicemail messages [14]. It was found that although the performance degrades significantly in the presence of transcription errors, it is possible to reliably determine the segments corresponding to phone numbers. The VoiSum system [15] proposed the generation and delivery of text summaries on mobile phone displays by extracting content words from the message transcriptions using a combination of lexical and prosodic features. Figure 7 depicts the MRROC curves generated in a binary extractive summarization task using lexical only, prosodic only and combination of lexical and prosodic features. Prosodic features as classifier inputs were found to help recall with cost in precision while combined lexical and prosodic features were up to 10% more robust than the combined lexical features alone, across all operating conditions.

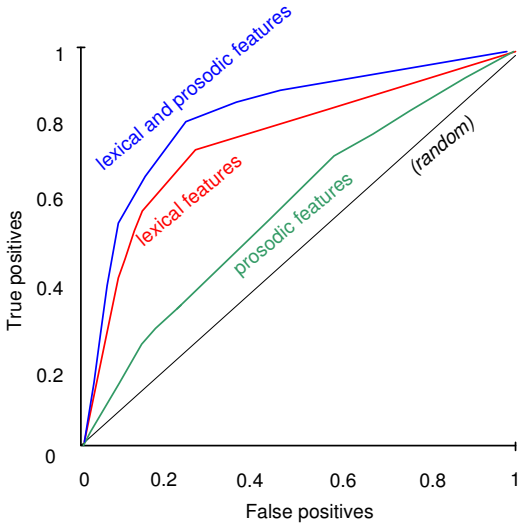

Figure 7. Maximum Realizable ROC curves using lexical only, prosodic only and combination of lexical and prosodic features in a voicemail extractive summarization task [15].

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). Language spoken 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 [33] has addressed some dialog-specific issues of summarization 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 [4]. Apart from the technical obstacles, a number of socio-cultural issues, such as privacy [10] are of higher importance in conversational speech rather than the other domains.

In human to human conversations, a great deal of information is conveyed by means other than speech and hence there are opportunities for synergy within user interfaces. Numerous and valuable content cues can be captured in video recordings (e.g., gestures, speaker localization). Such cues are the subject of several current

research projects whose goal is to extract relevant content from a variety of audio and visual sensor inputs and integrate it into a complete interaction index using statistical models [18, 23].

Future Directions

Content-based access to spoken audio 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. Traditionally, systems for content-based spoken audio access are built using spoken language processing and information retrieval components developed separately. 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. However, this fact has not yet been translated into a unified modeling approach, and as such the trainability and scalability of the component models remains limited. If this trend continues, there is a risk of failing to support very large spoken audio 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.

The providers of telecommunication and Web search services are expected to be the two main adopters of content-based to spoken audio technologies. On the one hand, as we are moving towards increased adoption of free but basic peer-to-peer calling based on the Internet protocol (IP), telecommunications companies will need to compete by offering value added services that depend on content-based access to spoken audio such as voicemail management, real-time language translation, recording and indexing of phone conversations. On the other hand, the major Web

search engine companies are eager to extend their offerings from the domain of hypertext to multimedia content including audio.

As of today Web search engines have not listed multimedia files mainly because of the technical difficulties in audio and video file search in comparison to hypertext files. As an intermediate step, it is possible to provide basic audio search services without performing content analysis. For instance, like current image and video search engines, one can perform basic audio search using the information found in file names. Another way would be to search through audio metadata that are already available (e.g., file headers), such as producer, length or date. Yet another possibility would be to exploit the associations between audio files. The latter approach would allow users to find similar audio files according to a number of attributes (e.g., topic, speaker, date, popularity, and types of background noise).

As users increasingly prefer to access content using handheld devices (smart phones and personal digital assistants), the associated application design implications of mobile access should be considered for content-based spoken audio access 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. news on demand) is required, user profiles that adapt over time and tend to reflect long-term information needs can be employed instead of repeated search 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 audio examples (e.g., related to a speaker or background conditions).

Continuous progress of the technologies reviewed will allow components that support content-based access to spoken audio to be integrated in numerous systems.

Examples of various potentially important applications that cover all levels (individual, business and government) are given in Table 1.

Table 1 Examples of potential applications based on content-based access of spoken audio.

Individual	• Personalized delivery of voicemail and news • Search of audio books and music • Analysis of personal audio recordings (meetings, presentations, telephone conversations)
Business	• Retrieval of help desk calls • Content management of corporate meetings
Government	• Access to audio proceedings (parliamentary sessions, court of law archives) • Access to cultural heritage archives • Monitoring unlawful conversations for security purposes

In order for applications such as the above to be successfully realized, research is needed in a number of areas. Given that relatively satisfactory speech recognition performance is now feasible in a number of domains, other less explored tasks (topic segmentation/ tracking, speaker detection/tracking, and summarization) need to be revisited. These 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. At the same time, integration of cues from image and video processing 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. Research in the above, will accelerate the transition from content access to content understanding.

Acknowledgements

We wish to thank Ed Schofield (ftw.), Miguel Carreira-Perpinan (OGI) and the two anonymous reviewers for their valuable comments on earlier drafts of this article. Author K. Koumpis was supported by a Marie Curie Fellowship.

Konstantinos Koumpis (koumpis@ftw.at) is a Senior Researcher at Vienna Telecommunications Research Center. Prior to this he was a Research Engineer at Domain Dynamics, a Research Associate at the University of Sheffield and a DSP Design Engineer at Nokia. He holds a DiplEng from the Technical University of Crete, an MSc from the University of Sussex and a PhD from the University of Sheffield. His current research is in the areas of mobile multimedia, networked embedded systems and innovation management.

Steve Renals (s.renals@ed.ac.uk) is Professor of Speech Technology in the School of Informatics at the University of Edinburgh, where he is also Director of the Centre for Speech Technology Research. Prior to this he was Reader in Computer Science at the University of Sheffield and held postdoctoral fellowships at the International Computer Science Institute, Berkeley and at the University of Cambridge. He holds a BSc from the University of Sheffield, and an MSc and a PhD from the University of Edinburgh. His research is in the areas of speech recognition, information access from spoken audio and models for multimodal data.

References

[1] D. Beeferman, A. Berger and J. Lafferty, "Statistical models for text segmentation," *Machine Learning,* Vol. 34, No. 1-3, pp. 177-210 1999.

[2] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. "Cross-task portability of a broadcast news speech recognition system," *Speech Communication*, Vol. 38, No. 3-4, pp. 335-347, 2002.

[3] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name," *Machine Learning*, Vol. 34, No. 1-3, pp. 211-231, 1999.

[4] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. on Speech and Audio Processing*, Vol. 12, No. 4, pp. 420-435, 2004.

[5] J. P. Campbell, "Speaker recognition: a tutorial," *Proc. of IEEE* , Vol. 85, No. 9, pp. 1437-1462, 1997.

[6] S. Dharanipragada, M. Frantz, J. S. McCarley and K. Papineni, "Statistical models for topic segmentation," *Proc ICSLP,* Beijing, 2000.

[7] G. Doddington "Speaker recognition based on ideolectical differences between speakers," In *Proc. Eurospeech*, pp. 2521-2524, Aalborg, Denmark, 2001.

[8] J. Garofolo, J. Lard, and E. Voorhees, "TREC-9 spoken document retrieval track: overview and results," In *Proc. 9th Text Retrieval Conf. (TREC-9)*, Gaithersburg, MD, USA, 2001.

[9] J. L. Gauvain and L. Lamel, "Large vocabulary continuous speech recognition: advances and applications" In *Proc. of IEEE*, Vol. 88, pp. 1181-1200, 2000.

[10] 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*, 2004.

[11] M. Hearst, "TextTiling: Segmenting text into multi-paragraph sub-topic passages," *Computational Linguistics,* Vol. 23, No. 1, pp. 33-64, 1997.

[12] J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker and G. Zamchick, "SCANMail: Browsing and searching speech data by content," In *Proc. Eurospeech*, pp. 1299-1302, Aalborg, Denmark, 2001.

[13] C. Hori, S. Furui, P. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," In *Proc. IEEE ICASSP*, Vol. 1, pp. 9-12, Orlando, FL, USA, 2002.

[14] J. Huang, G. Zweig, and M. Padmanabhan, "Information extraction from voicemail," In *Proc. of 39th Annual Meeting of Assoc. for Computational Linguistics*, Toulouse, France, 2001.

[15] K. Koumpis, and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, 2005, to be published.

[16] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, 1998.

[17] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz and A. Srivastava, "Speech and language technologies for audio indexing and retrieval". *Proc. IEEE*, Vol. 88, No. 8, pp. 1338–1353, 2000.

[18] I. McCowan, D. Gatica-Perez, S. Bengio, D. Moore, and H. Bourlard, "Towards computer understanding of human interactions". In *Proc. European Symposium on Ambient Intelligence (EUSAI)*, Eindhoven, The Netherlands, 2003.

[19] N. Morgan, Q. Zhu, A. Stolcke, K. Sönmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Gelbart, D. Ellis, G. Doddington, B. Chen, Ö. Çetin, Hervé Bourlard, M. Athineos, "Pushing the envelope – aside: Beyond the spectral envelope as the fundamental representation for speech recognition", This issue.

[20] J. M. Ponte, and W. B. Croft, "A language modeling approach to information retrieval," In *Proc. ACM SIGIR*, pp. 275-281, Melbourne, Australia, 1998.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition". In *Proc. of IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[22] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, Vol. 32, pp. 5-20, 2000.

[23] S. Renals and D. Ellis "Audio information access from meeting rooms," In *Proc. IEEE ICASSP*, Hong Kong, China, Vol. 4, pp. 744-747, 2003.

[24] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," *Proc. IEEE,* Vol 88, No 8, pp. 1270-1278.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management: an International Journal*, Vol. 24, No. 5, pp. 513-523, 1988.

[26] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol. 34, pp. 1-47, 2002

[27] E. Shriberg, A. Stolcke, D. Hakkani-Tür and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, Vol. 32, No. 1-2, pp. 127-154, 2000.

[28] E. Shriberg. "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, Vol. 31, No. 1, pp.153-169, 2001.

[29] J.-M. V. Thong, P. J. Moreno, B. Logan, B. Fidler, K. Maffey, and M. Moores, "Speechbot:an experimental speech-based search engine for multimedia content on the web," *IEEE Trans. Multimedia*, Vol. 4, No. 1, pp. 88–96, 2002.

[30] A. Verma and A. Kumar, "Techniques for voice conversion," This issue.

[31] T. Wakao, R. Gaizauskas, and Y. Wilks, "Evaluation of an algorithm for the recognition and classification of proper names," in *Proc. 16th Intl. Conf. on Computational Linguistics (COLING96)*, pp. 418-423, 1996.

[32] P. D. Wellner, M. Flynn and M. Guillemot, "Browsing Recordings of Multi-party interactions in ambient intelligence environments," In *Proc. CHI Workshop Lost in Ambient Intelligence?*, Vienna, Austria, 2004.

[33] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, Vol. 28, No. 4, pp. 447-485, 2002.