REAL-WORLD AUDIO INDEXING SYSTEMS

Beth Logan, Dave Goddeau and JM Van Thong

Hewlett-Packard Laboratories
One Cambridge Center
Cambridge MA 02142, U.S.A.
{beth.logan, dave.goddeau, jm.vanthong}@hp.com

ABSTRACT

We motivate the need for and describe the key components of real world audio indexing systems. In particular, we discuss the various flavors of such systems, the advantages and disadvantages of each, user interfaces, system architectures and evaluation issues. Throughout the paper, we give examples from our own experience of audio indexing using *SpeechBot* and its successor *NewsTuner*.

1. INTRODUCTION

The ever increasing amount of streaming and archived audio content on the web has the potential to provide a wealth of information and entertainment to anybody with internet access. Too often however, the user experience of wading through archives and live streams looking for interesting programs is a daunting task. As users move from a broadcast model of receiving content to TV and radio on demand, the need to search and browse live and archived programs will only increase.

Traditional search engines index and retrieve documents based on text representation. Other media, such as images, are indexed based on their surrounding text. This method may also be applied to index radio archives and streams. However, most audio streams are typically several hours in length and the surrounding text is often very limited. To search within programs or to find similar programs it is necessary to use metadata derived from the content itself or metadata created at production time, such as titles, abstracts, categories, close captions, and air dates. However, for radio and video without closed captions other content-based analysis and indexing techniques are needed.

Content-based indexing has been the focus of several research groups (e.g. Informedia [1], and [2, 3, 4, 5, 6]) and commercial systems are now available. Most systems combine one or more automatic content analysis techniques for segmenting and annotating the documents with annotations generated at production time.

2. CONTENT-BASED INDEXING SYSTEMS

In this section, we describe the principal techniques for content-based indexing of spoken audio. Audio document search performance is a trade-off between retrieval efficiency, vocabulary limitations, and query response time. We discuss the advantages and problems of two common approaches: word-based indexing, and subword-based indexing.

2.1. Word-based indexing

One approach to content-based audio indexing consists of automatically generating a transcription using a large vocabulary speech recognition system then using Information Retrieval (IR) algorithms to index the resulting textual documents. The index can then be used to retrieve relevant portions of the audio documents using standard word query terms. This has the advantage of being able to leverage the many years of work on scalable text indexing which has resulted in search engines capable of indexing the entire web. An example of such a system is shown in Figure 1.

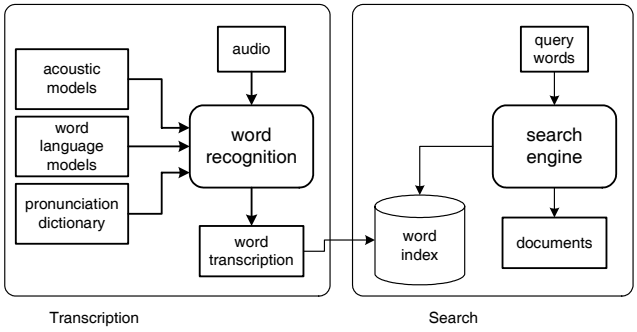

Fig. 1. Word-based indexing system.

There are two main problems with word-based audio indexing: recognition errors and vocabulary limitations. Audio content typically contains a variety of speakers and signal conditions, from studio announcers reading from scripts to audience members calling in on cell phones, to reporters transmitting live from the field. In addition, there are often overlays of noise and music which make the recognition task more challenging. Typical speech recognition error rates can vary from 5% for clean studio speech to as much as 50% for very noisy or conversational speech. The effect of recognition errors is mitigated by several factors. First, query terms are often longer words which tend to be recognized correctly. Second, poor acoustic conditions can often be compensated for by training the recognizer on the expected conditions. Finally, and probably most significantly, query words, particularly names and places, are often repeated several times in audio programs so there is a good chance that at least one occurrence will be correctly recognized, allowing the relevant program to be returned.

For query terms in the dictionary of the speech recognition system, good precision and recall can be achieved even for relatively high error rates, as demonstrated by the Hewlett Packard's

SpeechBot system [5, 7, 8]. It has been shown that up to 25% word error rate, search engine performance is equivalent to those running on exact text transcription [9]. Beyond that limit, retrieval accuracy begins to degrade.

2.2. The Out-Of-Vocabulary problem

Another serious problem for word-based audio indexing is Out Of Vocabulary (OOV) words in the audio documents or the queries. Typically the dictionary and language model are of insufficient size to cover all spoken documents and queries over time. For example, we have found that the number of unique words in *The New York Times* over a two year period is around 650,000 words, far more than the typical large vocabulary speech recognition dictionary. For most audio content, new words appear at a more-or-less constant rate, so expanding the dictionary to cover all previously seen words does not solve the problem, and can even reduce overall recognition accuracy by introducing more acoustically confusable words. A compromise solution is to update a fixed-size dictionary and language model over time to keep it up-to-date with current topics (assuming a suitable source of text data can be found). While, this can be effective in reducing the OOV rate in content recognition, there remains the problem of OOV words in queries. In the query space, we have found that over 10% of user queries to the *SpeechBot* system are OOV [7]. This problem has led to alternatives to word-based recognition approaches for audio indexing.

2.3. Subword-based indexing and searching

An alternative to word transcription is to use an intermediate subword representation for searching or indexing. A number of subword representations have been studied such as phonemes, phoneme sequences or syllables (e.g. [10, 11, 12]). The main advantage of these techniques is vocabulary independence; the recognizer dictionary uses a fixed, language specific set of units which can be composed to form any word in the language. The difficulty is that sub-word unit recognition error rates are much higher than word error rates, because of the inherent confusability of the short units, and the lack of language model constraints (which are much more effective at the word level). Therefore, sub-unit based indexing schemes must deal with this problem, either by carrying forward the recognition uncertainty in the form or a lattice of possibilities, or using approximate match techniques in the retrieval phase.

In the simplest case, that of phonemes, the audio is pre-processed to produce a *phoneme lattice*. This encodes multiple phoneme-sequence hypotheses. Each query is then decomposed into one or more phoneme sequences using a pronunciation dictionary or letter to sound rules. These strings are then searched for in the lattice returning exact or approximate matches. An example of such a system is shown in Figure 2.

Although such systems can improve recall by avoiding OOV problems, they typically have high false positive rates, reducing precision. Another disadvantage of this approach is that each query requires a time-consuming search through a lattice of alternative possibilities. This problem is can be addressed by indexing phoneme sequences [12, 13]. A related approach is to index words but convert the queries to a set of in-vocabulary words using letter to sound rules and a phoneme confusion matrix [14].

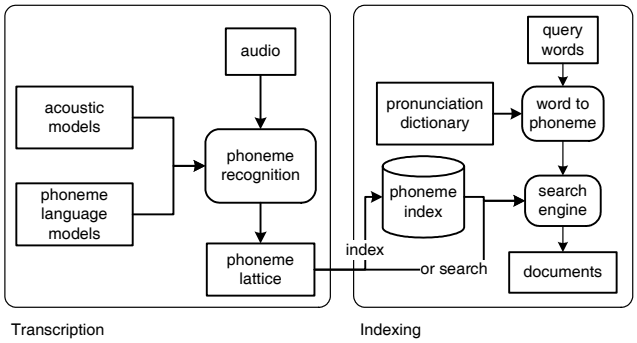

Fig. 2. Phoneme-based indexing system.

2.4. IR-motivated approaches to the OOV problem

Another way to attack the OOV problem is to use word-based indexing but try techniques from the IR community to improve retrieval. For example, query expansion and stemming have been found to be useful [15]. Another approach transforms the document and query to vectors describing their semantic content using techniques such as Latent Semantic Indexing (LSI) [16]. The queries and documents can then be compared in this semantic space where it does not matter if the query is OOV.

Finally, several studies (e.g. [17, 18, 19, 20]) have examined the combination of indexes, such as phonetic and word indexes, and have yielded promising results. It is difficult however to formulate combination rules that apply over all queries.

3. USER INTERFACE

Once annotated, the media can be presented to the end users according to their needs with possibly very different user interfaces. Annotations, or *metadata*, can be manually created, or automatically generated by content analysis systems, as described below. In the following, we define metadata as any form of annotation of the content, time coded or not, e.g. title, summary, word transcription, or segment boundaries.

3.1. Topic hierarchies

When the documents are classified and grouped into categories, the user can browse topic hierarchies. Categories are typically chosen by content producers with the objective of offering suggestions for related material. At best, this may help users discover new stories of interest. In practice however, a rigid and static taxonomy proves to be an ineffective means of grouping similar content for browsing. First, producers may disagree on how to categorize stories, leading to inconsistency in the data. Second, users may be confused by topic headings and unable to find the documents they were looking for. Finally, a fixed hierarchy may not be flexible enough to respond to emerging events, even with updates and additions. From a user perspective, traversing topic hierarchies may be tedious, time consuming and ultimately frustrating.

3.2. Keyword-based search engines

User behavior observed in focus groups indicates that the majority of people bypasses navigation and instead prefer keyword

searches. *SpeechBot* is an example of an audio search engine, which implements such a keyword search user interface [8]. When presented with a query, the engine returns a list of matching programs and the locations of those matches within individual programs. The time-coded transcription computed by the system allows random access within the audio document. The indexer employs a modified version of the query engine developed for the AltaVista search service. To sort the documents by relevance, we use a term frequency / inverse document frequency *(tf.idf)* IR metric [21] augmented by information about proximity of the query terms in the transcription. The further apart the terms are the lower the score. The current implementation indexes over 15,000 radio programs.

3.3. Semantic search

In collaboration with WBUR[1], we have recently released *NewsTuner* [22], an alternate user interface to more efficiently access archived material [23]. NewsTuner is a player which combines live broadcasts, archived audio, chat, and studio cameras into a single application. Since searching and browsing are not mutually exclusive, *NewsTuner* offers two ways for users to find an audio file: keyword and similarity searches. Keyword search allows users to hunt through transcripts or production metadata of a collection of audio stories looking for an exact word match. Similarity search returns a list of audio files that contain semantically-similar content to a selected story. We use Hofmann's PLSA [24] on producer generated summaries to define similarity. The same technique can be used on inaccurate transcripts, as demonstrated in [25]. Keyword searching is effective for users who want to find a specific piece of content, while similarity matching is good for users browsing from topic to topic.

4. SYSTEM ARCHITECTURES FOR AUTOMATIC CONTENT PROCESSING

The size of media archives and the expensive computations needed to process the content require dedicated scalable system architectures.

Real-time systems adopt pipeline architectures: the incoming audio/video signal is digitized, and the output is fed into a sequence of automatic content processing components that generate time-coded metadata. The output is an annotated stream that can then be indexed. However these solutions usually require each component to be implemented as a real-time streaming process, possibly imposing trade-offs on accuracy of the created metadata.

To process pre-recorded material such as that from existing archives, batch systems can be used (e.g. [26]). The real-time streaming constraint is no longer an issue, so the task can be parallelized across several servers. The media stream is split into short segments and each segment is analyzed separately. The segments should overlap to avoid edge effects. Since the produced metadata is time-coded, the resulting chunks of the different processes can be easily recomposed into a seamless data stream. These approaches are well suited for long running, non-real-time algorithms.

Hybrid systems can take advantage of both approaches by exploiting data parallelism within a streaming architecture. A directed acyclic graph (DAG) describes how the analysis components are connected to each other; one output of one module being the input of the next as in a producer-consumer model. Each non-streaming component can be replicated and executed on different data segments. With enough computing resources, real-time throughput with bounded latency can be achieved.

[1]WBUR, a National Public Radio (NPR) affiliated radio, is based in Boston, MA, online at *www.wbur.org*

5. EVALUATION METHODS

An established way to compare retrieval systems is to report mean 11-pt average precision (e.g. [21]). This is an estimate of the area under the precision *vs.* recall curve averaged over all queries. An ideal system has a mean average precision of 1.0.

While such a metric is a good measure of performance, it assumes that recall can be computed. For annotated data, such as that provided by the TREC Spoken Document Retrieval (SDR) track [27], this is not a problem. For systems indexing "found" data on the web, the number and location of relevant hits for a typical query is unlikely to be known or easily determined. Thus researchers typically fall back on reporting average top 5 and top 10 precision since this indicates how many relevant results are found on average on the first few pages of returned results and is a good measure of the system effectiveness. As an example of performance, *SpeechBot* has been shown to have an average top 5 precision of 65% [5] on a set of 40 in-vocabulary queries.

The queries used for evaluations should be selected carefully. Buckley et. al. [28] recommend that at least 25 and preferably 50 queries be used for an evaluation for which average precision is the metric. In our evaluations, we have used queries from our user logs as much as possible. For studies in which we wished to use unambiguous proper names and report recall, insufficient queries were found in the logs. We therefore augmented these queries with artificial queries derived from the true audio transcriptions. In such cases, we chose queries of 1, 2 or 3 words from the true audio transcriptions in similar proportions to that observed in the user logs. In general, real world queries are rarely greater than 3 words.

6. CONCLUSION

This paper has discussed the major approaches to indexing spoken audio content is the absence of pre-supplied transcriptions or metadata. Word-based approaches seem to offer the best precision and lowest retrieval cost, but deal poorly with OOV queries and content. Sub-word unit based indexing approaches avoid the OOV problem, but at the cost of reduced precision and increased computation for query processing. We believe that audio search performance can be improved by using fusion techniques, either at the transcription or indexing level. The ideal combination has yet to be found.

More optimistically, current recognition technology allows for the development of systems that usefully make audio content available for search and indexing. With the proper system architecture, this can be done it real-time making live content available for indexing and retrieval as it is produced. Finally, creative user interfaces can mitigate some of the inherent problems in audio indexing.

7. ACKNOWLEDGMENTS

Pedro Moreno, now at Google, was an essential contributor to the media indexing projects done at Digital-Compaq-HP CRL. Radio program indexing experiments on a large scale has been made possible at HP thanks to the collaboration with WBUR, the Boston local NPR affiliated radio. We'd particularly like to recognize Jon Marston and Gavin MacCarthy of WBUR who played an integral part in the design and implementation of the WBUR *NewsTuner*.

8. REFERENCES

[1] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock, "Informedia: News-on-demand experiments in speech recognition," in *ARPA Speech Recognition Workshop*, 1996.

[2] S. E. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. C. Woodland, "The Cambridge University spoken document retrieval system," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[3] D. Abberley, G. Cook, S. Renals, and T. Robinson, "Retrieval of broadcast news documents with the THISL system," in *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 1999, pp. 128–137.

[4] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Nielser, A. Tuerk, and S. J. Young, "Experiments in broadcast news transcription," in *Proc. ICASSP*, May 1998.

[5] JM. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain, "Speechbot: a speech recognition based audio indexing system for the web," in *Proc. International Conference on Computer-Assisted Information Retrieval (RIAO)*, 2000.

[6] M. Clements, P. S. Cardillo, and M. S. Miller, "Phonetic searching vs. LVCSR: How to find what you really want in audio archives," in *20th Annual AVIOS Conference*, 2001.

[7] B. Logan, P. Moreno, JM. Van Thong, and E. Whittaker, "An experimental study of an audio indexing system for the Web," in *Proc. ICSLP*, 2000.

[8] "Speechbot web site," Dec. 1999, www.speechbot.com.

[9] A. Hauptmann, R. Jones, K. Seymore, S. Slattery, M. Witbrock, and M. Siegler, "Experiments in information retrieval from spoken documents," in *Broadcast News Transcription and Understanding Worshop*, 1998, pp. 175–181.

[10] D. A. James, "A system for unrestricted topic retrieval from radio news broadcasts," in *Proc. ICASSP*, 1994.

[11] P. Schaeuble and M. Wechsler, "First experiences with a system for content based retrieval of information from speech recordings," in *IJCAI-95*, 1995.

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

[13] S. Dharanipragada and S. Roukos, "A multistage algorithm for spotting new words in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 542–550, Nov. 2002.

[14] B. Logan and JM. Van Thong, "Confusion- based query expansion for oov words in spoken document retrieval," in *Proc. ICSLP*, 2002.

[15] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. Spark Jones, "Effects of out of vocabulary words in spoken document retrieval," in *ACM SIGIR conference on research and development in information retrieval*, July 2000, pp. 372–374.

[16] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *American Society for Information Science*, vol. 6, no. 41, pp. 391–407, 1990.

[17] G. J. F. Jones, J. T. Foote, K. Spark Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," in *ACM SIGIR*, 1996, pp. 30–38.

[18] K. Ng and V. Zue, "Information fusion for spoken document retrieval," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 2405–2408.

[19] B. Logan, Pedro Moreno, and Om Deshmukh, "Word and sub-word indexing approaches for retuding the effects of OOV queries on spoken audio," in *HLT*, 2002.

[20] B. Logan, P. Prasangsit, and P. Moreno, "Fusion of semantic and acoustic approaches for spoken document retrieval," in *ICSA Workshop on Multilingual Spoken Document Retrieval*, 2003.

[21] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[22] "Newstuner web site," May 2004, www.newstuner.org.

[23] J. Marston, G. MacCarthy, B. Logan, P. Moreno, and JM Van Thong, "News tuner: A simple interface for searching and browsing radio archives," in *ICME*, 2004.

[24] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR1999*, 1999.

[25] D. Blei and P. J. Moreno, "Topic segmentation with an aspect hidden markov model," in *SIGIR2001*, 2001.

[26] H. Mandviwala, S. Blackwell, C. Weikart, and JM. Van Thong, "Multimedia content analysis and indexing: Evaluation of a distributed and scalable architecture," in *SPIE's International Symposium on ITCom 2003*, 2003.

[27] J. Garfolo, E. Vorhees, C. Auzanne, V. Stanford, and B. Lund, "Spoken document retrieval track overview and results," in *Proceedings of the 7th Text Retrieval Conference (TREC-7)*, 1998.

[28] C. Buckley and E. Voorhees, "Evaluating evaluation measure stability," in *ACM SIGIR*, 2000, pp. 33–40.