Spoken Content-Based Audio Navigation (SCAN)

John Choi, Donald Hindle, Julia Hirschberg, Fernando Pereira, Amit Singhal and Steve Whittaker

AT&T Labs – Research, Florham Park, New Jersey, USA
{hindle/julia/pereira/singhal/stevew}@research.att.com

ABSTRACT

We describe SCAN, a system for retrieving and browsing speech documents from large audio corpora that uses new information retrieval and speech processing techniques to create easily navigable presentations of documents relevant to a user query. Experiments show that the new interface is more effective than simple speech-alone interfaces.

1. INTRODUCTION

The goal of this research is to provide intelligent interactive access to the increasing quantity of public (broadcasts and Web recordings) and private (voicemail, meeting records) stored speech. To achieve this goal, we are investigating methods for using the structure and content of spoken documents for more effective retrieval and browsing. In particular, we exploit intonational structure to segment spoken documents in more easily retrievable and browsable units, and we use parallel text corpora to improve the language models used in automatically transcribing spoken documents and to increase the effectiveness of information retrieval from the resulting transcriptions.

The first instantiation of our ideas is SCAN (Spoken Content-based Audio Navigator), which operates on the NIST TREC-6/7 SDR corpora[1] Like other TREC SDR systems, it uses automatic speech recognition techniques enhanced by some preliminary acoustic channel and intonational analysis to produce an errorful transcription of the news broadcasts, which have been segmented into stories by NIST labelers for the retrieval task. Stories relevant to an input text query are retrieved by a modified version of the SMART information retrieval system [1]. Results of the recognition and retrieval stages are then passed to a graphical user interface, which presents them to the user in several different ways, to support better browsing and navigation. Below we describe the current state of our system, updating previous descriptions [2], and briefly describe some results of our system evaluations.

2. THE *SCAN* SYSTEM

2.1. Paratone Detection

The news stories in our corpus may be up to 25 minutes long. To segment these for our speech recognizer, as well as for our end users, we trained a decision tree to recognize intonational phrase boundaries, which can then be merged into intonational paragraphs, or *paratones* [4]. The classifier was trained on the Boston Directions Corpus [3], which had previously been hand labeled for intonational boundaries using the ToBI labeling conventions [6]. Acoustic features which best predicted intonational boundaries in this data included fundamental frequency (F0), RMS energy, and autocorrelation peaks, and were derived from the Entropic WAVES pitch-tracker `get_f0`. The best-performing classifiers on the BDC corpus performed at precision/recall rates of 92,95%/74,71% on a hand-labeled test set from the TREC SDR corpus (230 sec of an NPR broadcast containing 88 intonational phrases).

The classifier is used to segment the speech stream for the recognizer into 'chunks' around 20 msec long, by locating the closest intonational phrase boundary to this limit. We believe this is preferable to using fixed-size units, which can begin or end within words, or break apart words which should be considered together in the language model. Currently, these same units are used for visual browsing and play-back in the SCAN GUI. Better choice of boundaries for these paratones can be made using simple pausal duration information, for a given speaker; that is, longer pauses are reliably correlated with topic beginnings. However, it is difficult to find topic boundaries reliably across speakers, due to differences in speaking rate.

2.2. Classifying Channel Conditions

The intonational paratones classified by CART are then passed to another classifier, which divides them into wideband or narrowband speech. The TREC SDR data is labeled more specifically as to recording conditions, including information about background noise and music; however, previous experiments showed that a simple wide or narrow band distinction performed as well for recognition purposes as a more complex set of distinctions.

The channel classifier is based on full covariance Gaussian mixture models, initialzed using vector quantization, and trained using the EM algorithm. It uses 31-element mel-scaled filter-bank coefficients (dB) as input. The channel-classified paratone units are then passed to an automatic speech recognizer, which will use the channel hypothesis proposed by the classifier to determine which of two sets of acoustic models to apply to this segment.

2.3. The Speech Recognizer

Our recognizer is based on a standard time-synchronous beam search algorithm. The probabilities defining the transduction from text-dependent phone sequences to word sequences are estimated on word level grapheme-to-phone mappings and are implemented

[1] A subset of the DARPA HUB-4 Broadcast News corpus, which includes news broadcasts from the major networks and CNN.

in the general framework of weighted finite-state transducers [5]. Transducer composition is used to generate word lattice output.

We use continuous density, three-state, left-to-right, context-dependent hidden Markov phone models. These models were trained on 39-dimensional feature vectors consisting of the first 13 mel-frequency cepstral coefficients and their first and second time derivatives. Training iterations included eigenvector rotations, k-means clustering, maximum likelihood normalization of means and variances and Viterbi alignment. The output probability distributions consist of a weighted mixture of at most 12 Gaussian components with diagonal covariance. As explained above, the training data were divided into wideband and narrowband partitions, resulting in two acoustic models.

Language Models We use a two pass recognition process. In the first pass, we build word lattices for all the speech, using a minimal trigram language model and a beam determined heuristically to provide word lattices of manageable size. In the second pass, these word lattices are rescored by removing the trigram grammar weights while retaining the acoustic weights and then composing the resulting lattices with a 4-gram language model. The best path is extracted from the rescored lattices.

Both the first pass trigram language model and the rescoring 4-gram model are standard Katz backoff models, using the same 237,000 word vocabulary. For choosing the vocabulary, all of the words from the SDR98 training transcript were used. This base vocabulary was supplemented with all words of frequency greater than two appearing in the New York Times and LA Times segments of LDC's North American News corpus (LDC Catalog Number: LDC95T21, see `www.ldc.upenn.edu`), in the period from June 1997 through January 1998. The vocabulary includes about 5,000 common acronyms (e.g "N.P.R."); language model training texts were preprocessed to include these acronyms.

Language model training was based on three transcription sources (the SDR98 training transcripts, HUB4 transcripts, transcripts of NBC Nightly News) and one print source (the LDC NA News corpus of newspaper text). The first-pass trigram model was built by first constructing a backoff language model from the 271 million words of training text, yielding 15.8 million 2-grams and 22.4 million 3-grams. This model was reduced in size, using the approach of Seymore and Rosenfeld [7], to 1.4 million 2-grams and 1.1 million 3-grams. When composed with the lexicon, this smaller trigram model yielded a manageable sized network. The second pass model used 6.2 million 2-grams, 7.8 million 3-grams, and 4.0 million 4-grams. For this model, the three transcription sources (SDR, HUB4, NBC) were in effect interpolated with the text source (NA News), with the latter being given a relative weight of 0.1.

The performance of our recognition component on the TREC7 test set was 32.4% word error rate (WER). This was slightly better than the 'medium error' transcriptions provided by NIST in the TREC7 competition, although considerably worse than the 24.8% WER of the top recognizer on this test set. Despite this handicap, our retrieval results were quite good, due to some innovations in expanding both the queries and the documents in our collection.

2.4. The Information Retrieval System

We use a modified version of the SMART information retrieval engine to perform audio 'document' retrieval from automatic transcriptions. SMART is based on the vector space model of information retrieval. It generates weighted term (word) vectors for the automatic transcriptions of the documents (where each document is a story as marked by the NIST labelers). User queries are also converted into weighted term vectors. Vector inner-product similarity computation is then used to rank documents in decreasing order of their similarity to the user query. SMART preprocesses the automatic transcriptions of each news story by tokenizing the text into words, removing common words that appear on its stop-list, and performing stemming on the remaining words to derive a set of terms. Our version of SMART also uses statistically-selected phrases tailored to our corpus.

User queries are typically short, and enriching such short queries with words related to the query (*query expansion*) is a well-established technique for improving the retrieval effectiveness. For example, if the user query is 'find reports of fatal air crashes,' adding words like *flight*, *airline*, or *safety*, to the query yields better retrieval results. We use query expansion in SCAN, as described in [8]. In brief, the initial user query is first used to locate some top-ranked documents that are well-related to the user query. Words that are frequent in those documents are then added to the query.

We also perform *document expansion*, to compensate for some of the recognizer's mistakes, adding words that "could have been present" to our automatic transcriptions of each news story. In a set of experiments on doing document expansion from the NA News corpus alone, we found that when speech stories were *not* reported in the print media, document expansion hurt retrieval performance, since unrelated words were added to the story. To contain this problem, we force the expansion algorithm to choose only those words that are also present in the word lattice generated by our recognizer for the speech story. This restriction guarantees that the words being added to a document are also proposed by the speech recognizer, albeit with a low confidence. Document expansion is thus performed by first taking the one-best recognition output for a given story and using that as a query itself on the larger NA News corpus. From the documents retrieved, we identify those terms that appear in the recognition word lattice from which our one-best output was derived, and add the top 25% (up to 50) new terms occuring in at least half the top 20 retrieved documents to the transcription of that story. These parameters were chosen somewhat arbitrarily, based on a quick inspection of the expansion terms and our experience with relevance feedback. We had no testbed to tune our parameters.

We tested the retrieval effectiveness of SCAN on TREC-7 SDR track data [9]. Results show that when retrieval is done on automatic transcriptions, average precision is 0.4371, just 3.9% behind retrieval from perfect transcriptions. Document expansion removes this difference and retrieval from expanded documents is at par with retrieval from human transcriptions, at 0.4535. Query expansion improves the retrieval effectiveness for all transcriptions. The average precision for retrieval from human transcriptions improves to 0.5083. The gains for retrieval from expanded documents are stronger, and the average precision improves to 0.5300 — actually surpassing retrieval from human transcriptions (0.5083) by 4.3%. These results indicate that doing information retrieval from spoken documents using automatic transcriptions is quite feasible. The retrieval effectiveness of SCAN's retrieval component is at par with doing retrieval from human transcriptions.

2.5. The User Interface

Information retrieval from audio data is sharply different from information retrieval from text, due to the linear nature of speech, and the differences in human capabilities for processing speech versus text. Yet, to date, user interfaces for both types of retrieval systems have focused on search, where the goal is simply to identify a ranked set of text or audio documents relevant to the user's query. For detailed information seeking in textual material, users can easily visually scan and browse the retrieved texts to identify relevant regions. In a speech corpus, however, it is apparent that user interfaces providing only document level search are insufficient: a story in the NIST Broadcast News corpus, for example, can be as long as 25 minutes. Given the sequential nature of speech, it is extremely laborious to scan through multiple long stories to obtain an overview of their contents, or to identify specific information of direct relevance within them. Interfaces for accessing speech archives therefore need to support local navigation within speech documents in addition to search.

3. AUDIO BROWSING STUDIES

To identify local browsing needs we conducted a series of empirical studies [11]. First we studied heavy users of a current major speech archiving technology, voicemail, to discover their needs and problems. Next, we compared two very simple speech browsers empirically, to understand user behavior in information retrieval tasks that involved finding specific information and summarizing larger chunks of information. From experienced audio browsing/retrieval users, we identified primary needs and difficulties with current technology. We learned that two major problems for users were *scanning*, that is, navigating to the correct message or relevant part of the message, and *information extraction*, that is, accessing specific facts from within a message. To address these problems, 72% of users usually took notes, either full transcriptions or simpler message indexing, abstracting only key points to be used later to locate the original message in the archive. In our laboratory studies, we found that subjects experienced serious problems with local navigation, even in a very small speech archive of short voicemail messages. They learned the global structure of the archive, but had trouble remembering individual message contents. Information extraction tasks were extremely hard, particularly when multiple facts needed retrieving, and users repeatedly replayed material they had just heard, suggesting problems with remembering local message structure.

3.1. WYSIAWYH Paradigm and empirical evaluation

As a result of these findings, we proposed a new paradigm for speech retrieval interfaces: "what you see is (almost) what you hear" (WYSIAWYH) [10]. This is a multimodal approach (see Fig 1) based on the notion of providing a *visual analog* to the underlying speech.

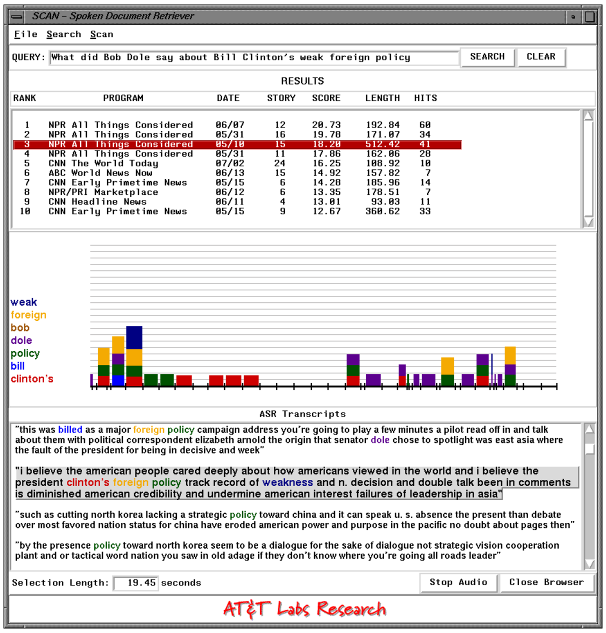

Figure 1: WYSIAWYH Browser

We use text formatting conventions (such as headers and paragraphs) to take advantage of well-understood text conventions to provide useful local context for speech browsing. The interface presents two types of visual information: an abstract overview and a formatted transcript. This visual information provides methods for users to index into the original speech documents. By depicting the abstract structure of an audio document in the overview, and by providing a formatted transcription, we hoped to make visual scanning and information extraction more effective, addressing the problems of local navigation identified in our user studies. We implemented this paradigm in a multimodal user interface to SCAN, described below, and evaluated our results in a comparison with a simple speech interface.

For each story, we make use of our (errorful) ASR transcription, paratone segmentation, SMART-selected query terms and their weightings, and SMART relevance-ranked documents. The SCAN UI has four components to access these:

- The search component provides rapid access to both the audio and transcripts of the set of potentially relevant documents. SMART retrieves these via a search panel at the top of the browser. Results are depicted in the Results panel immediately below, which presents a relevance-ranked list of 10 audio documents, with additional information, including program name and story number, date, relevance score, length (in seconds), and total hits (number of instances of query words)
- The visual overview component provides high-level information about individual audio documents, so users can

rapidly scan to locate potentially relevant regions. It displays which query terms appear in each paratone of the story. Each query word is color coded, and each paratone is represented by a column in a histogram. The width of the column represents the relative length of that paratone. The height of each column in the histogram represents the overall query word density (number of instances of the query terms normalized for the paratone length) within the paratone. Users can directly access the speech for any paratone by clicking on the corresponding column of the histogram.

- The automatic transcript supports information extraction, providing detailed, if sometimes inaccurate, information about the contents of a story. Query terms in the transcript are highlighted and color-coded, using the same coding scheme used in the overview panel. Users can play a given paratone by clicking on the corresponding paragraph in the transcript.
- A simple play bar represents a single story, which users can access randomly within the bar, plus start and stop buttons to control play for this component and others.

To test our hypotheses about the usefulness of our WYSIAWYH paradigm in supporting local browsing, we compared the SCAN browser, with a control interface that gave users only the search panel and the player component. Subjects performed three different types of tasks: relevance judgments for five stories retrieved for a query; finding simple facts; and summarization of a given story in 4-6 sentences. The experimental design was randomized within subjects. Twelve subjects were given 4 of each of the 3 tasks types. For half they used the SCAN browser, and the control browser for the other half. For each question we measured outcome information: time to solution and quality of solution (as assessed by two independent judges); collected process information (number, type, target story, and duration of browsing and play operations). We also collected subjective data, including subject ratings of task difficulty and the quality of the automatic transcript for the SCAN condition. We also encouraged subjects to "think aloud" as they carried out the tasks and gave them a post-test survey asking about relative task difficulty, how well the SCAN UI supported each task, overall browser quality , how the browser might be improved, quality of the transcript, and what led them to evaluate the transcript positively or negatively.

We found that users generally performed better with the SCAN WYSIAWYH browser than with the control, in terms of time to solution, solution quality, perceived task difficulty, and users' rating of browser usefulness. With the SCAN browser, people played much less speech, although they executed more play operations. We infer that the SCAN browser allowed users to play more selectively. However, while the SCAN UI improved performance in the fact-finding and relevance ranking tasks significantly, it did not improve the summarization task.

4. CONCLUSIONS AND FURTHER RESEARCH

SCAN currently provides a means of finding and browsing information in a large speech database. It has been shown to retrieve documents with high effectiveness. It also improves audio browsing in two important tasks, fact-finding and document relevance-ranking, compared with simple speech-only browsing. Next steps to improve both areas are to identify relevant regions within retrieved audio documents. Our system's failure to improve summarization we will address by providing users with automatic document summarization, document topic segmentation and document outlining as a first approximation which can be fleshed out by selective listening. We also intend to improve the paratone detector, by incorporating relative pausal duration between intonational phrases into the presentation of our ASR transcription, so that browsing can take advantage of inferred topic segmentation. Additional steps will involve taking our Broadcast News browsing beyond the NIST corpus to handle current material as it is broadcast. We are also porting our current technology to the voicemail domain.

5. REFERENCES

1. C. Buckley. Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, NY 14853, May 1985.
2. J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Chagnolleau, C. Nakatani, F. Pereira, A. Singhal, and S. Whittaker. Scan - speech content based audio navigator: A systems overview. In *Proceedings of ICSLP-98*, Sydney, 1998.
3. J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of ACL-96*, Santa Cruz, 1996.
4. J. Hirschberg and C. Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of ICSLP-98*, Sydney, 1998.
5. F. C. N. Pereira and M. D. Riley. Speech recognition by composition of weighted finite automata. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 431–453. MIT Press, Cambridge, Massachusetts, 1997.
6. J. Pitrelli, M. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of ICSLP-94*, Yokohama, 1994.
7. K. Seymore and R. Rosenfeld. Scalable backoff language models. In *Proceedings of the ICSLP-96*, 1996.
8. A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira. AT&T at TREC-7. In E. Voorhees and D. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).
9. E. Voorhees and D. Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In E. Voorhees and D. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).
10. S. Whittaker, J. Choi, J. Hirschberg, and C. Nakatani. "what you see is almost what you hear: Design principles for accessing speech archives. In *Proceedings of ICSLP-98*, Sydney, 1998.
11. S. Whittaker, J. Hirschberg, and C. Nakatani. Play it again: a study of the factors underlying speech browsing behavior. In *Proceedings of CHI '98*, Los Angeles, 1998.