FINDING INFORMATION IN AUDIO: A NEW PARADIGM FOR AUDIO BROWSING AND RETRIEVAL

Julia Hirschberg, Steve Whittaker, Don Hindle, Fernando Pereira, and Amit Singhal

AT&T Labs – Research
Shannon Laboratory, 180 Park Ave., Florham Park, NJ 07932, USA

ABSTRACT

Information retrieval from audio data is sharply different from information retrieval from text, not simply because speech recognition errors affect retrieval effectiveness, but more fundamentally because of the linear nature of speech, and of the differences in human capabilities for processing speech versus text. We describe SCAN, a prototype speech retrieval and browsing system that addresses these challenges of speech retrieval in an integrated way. On the retrieval side, we use novel document expansion techniques to improve retrieval from automatic transcription to a level competitive with retrieval from human transcription. Given these retrieval results, our graphical user interface, based on the novel WYSIAWYH ("What you see is almost what you hear") paradigm, infers text formatting such as paragraph boundaries and highlighted words from acoustic information and information retrieval term scores to help users navigate the errorful automatic transcription. This interface supports information extraction and relevance ranking demonstrably better than simple speech-alone interfaces, according to results of empirical studies.

1. INTRODUCTION

To date, user interfaces for both text and speech retrieval systems have focussed on search, where the goal is simply to identify a ranked set of text or audio documents relevant to the user's query. In text it may be that, for more detailed information seeking, users can easily scan and browse the retrieved texts to identify relevant regions. In a speech corpus, however, it is apparent that user interfaces providing only (audio) document retrieval are insufficient. For instance, a story in the NIST Broadcast News corpus 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. In addition to searching for relevant documents, interfaces for accessing speech archives therefore also need to support local navigation within such documents. Based on two user studies to identify current user problems and strategies for searching speech archives, we propose a new paradigm for multimodal user interfaces to speech data, and describe empirical evaluation of a system built according to this paradigm.

1.1. Initial Studies

To identify local browsing needs we conducted a series of empirical studies. First we studied heavy users of a current major speech archiving technology, voicemail, to discover their needs and problems [7, 19]. 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 [20, 10]. From our experienced audio browsing/retrieval users, we identified primary needs and difficulties with current technology. We learned that *scanning*, that is, navigating to the correct message or relevant part of the message, and *information extraction*, accessing specific facts from within the message presented major difficulties for users. 72% of users usually took notes, either full-transcription 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 could learn 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.

1.2. What You See Is (Almost) What You Hear

As a result of these findings, we proposed a new paradigm for speech retrieval interfaces: "what you see is (almost) what you hear" (WYSIAWYH) [18]. This is a multimodal approach (Figure 1) based on the notion of providing a *visual analog* to the underlying speech.

John Choi and Christine Nakatani made important contributions to the SCAN system and to the empirical evaluations described here.

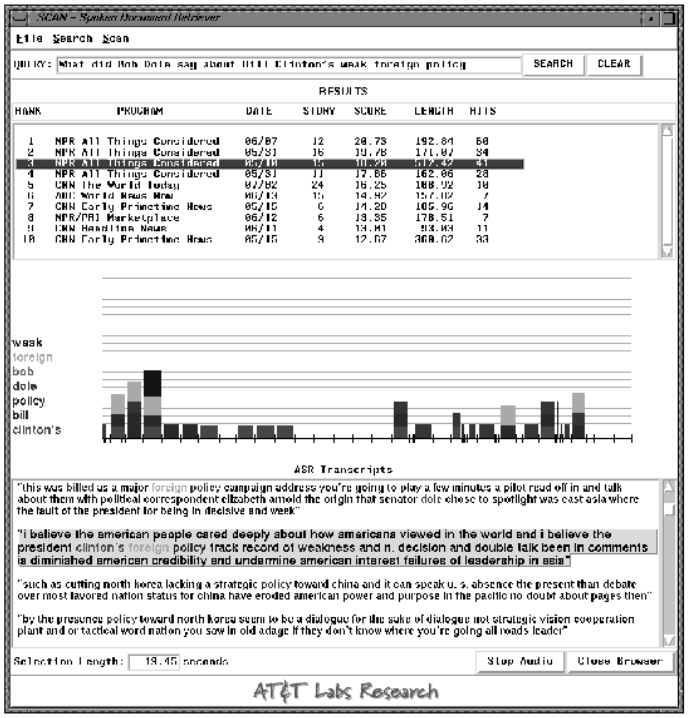

Figure 1: WYSIAWYH Browser

We use text formatting conventions (such as headers and paragraphs) to take advantage of well-understood text conventions and 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.

2. THE SCAN SYSTEM

SCAN operates by segmenting speech documents into *paratones*, or audio paragraphs, using acoustic information, classifying the recording conditions for each segment (narrowband or other) and employing automatic speech recognition (ASR) on each. We combine ASR results for each paratone so that for each audio document we have its corresponding (errorful) transcription. Terms in each transcript are then indexed for subsequent retrieval by an adaptation of the SMART information retrieval system [13].

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 CART [1] classifier to recognize intonational phrase boundaries, which can then be merged into intonational paragraphs, or *paratones* [6, 5]. The classifier was trained on the Boston Directions Corpus (BDC) [4], which had previously been hand labeled for intonational boundaries using the ToBI labeling conventions [15, 12]. 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 two best classifiers on the BDC corpus performed at precision/recall

rates of 0.92/0.74 and 0.95/0.71 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 [2, 3, 4]. 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 a simple Gaussian-mixture-based classifier that divides them into wide-band or narrow-band 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.

2.3. The Speech Recognizer

Our recognizer uses a standard time-synchronous beam search algorithm operating on a *weighted finite-state transducer* [11, 9] representing the context-dependency, lexical and language model constraints and statistics of the recognition task. Context-dependent phones are modeled with continuous density, three-state, left-to-right hidden Markov models. State densities are modeled by mixtures of up to 12 diagonal-covariance Gaussians over 39-dimensional feature vectors (first 13 mel-frequency cepstral coefficients and their first and second time derivatives).

2.3.1. Lexicon

We use a 237,000 word vocabulary including all the words in SDR98 training transcript, common words on newswire of the same time period, and 5,000 common acronyms.

2.3.2. 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 with a more detailed 4-gram language model. The best path is extracted from the rescored lattices. Both models are based on the Katz back-off technique [8] and are pruned using the shrinking method of Seymore and Rosenfeld [14].

2.3.3. ASR Performance

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 system [13] to perform audio `document' retrieval from automatic transcriptions. In SMART, both documents and queries are represented as term-indexed weight vectors, and documents retrieved for a query are ranked according to the inner product of the query and document vectors.

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 retrieval effectiveness [16]. In brief, the initial user query is first used to locate some top-ranked documents that are related to the user query, and 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 the automatic transcription of each news story. We first take the one-best recognition output for a given story and use that as a query itself on a larger text 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 occurring in at least half the top 20 retrieved documents to the transcription of that story. The process is described in detail elsewhere [16].

We tested the retrieval effectiveness of SCAN on TREC-7 SDR track data [17]. 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%.

3. THE USER INTERFACE

For each story, we make use of the (errorful) ASR transcription, paratone segmentation, SMART-selected query terms and their weightings, and SMART relevance-ranked documents. The SCAN UI (Figure 1) 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 shows the query terms that 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.

4. EMPIRICAL EVALUATION

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 task 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. Subjects were encouraged to "think aloud" as they carried out the tasks and answered 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

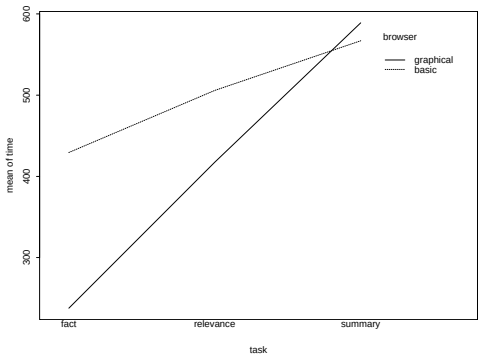

Figure 2: Solution Time for Each Task

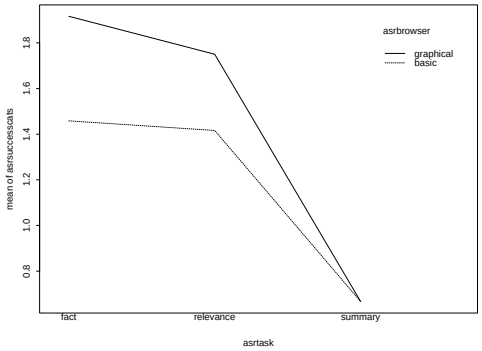

Figure 3: Solution Quality for Each Task

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 (as shown in Figures 2 and 3).

5. CONCLUSION AND FURTHER RESEARCH

SCAN currently provides a means of finding and browsing information in a large speech database. It has been demonstrated to retrieve documents with high effectiveness. It also improves audio browsing in two important tasks, fact-finding and document relevance-ranking, when compared with simple speech-only browsing. Next steps to improve both areas are to identify relevant regions within retrieved audio documents.

Our SCAN GUI does not appear to improve summarization. We believe that automatic speech summarization, document topic segmentation and document outlining may be important techniques to aid in audio document summarization by providing a first approximation which users can then flesh 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 also plan to apply techniques developed for news stories to a voicemail domain; both projects are currently underway.

6. REFERENCES

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA, 1984.

[2] Barbara Grosz and Julia Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October 1992. ICSLP.

[3] Julia Hirschberg and Barbara Grosz. Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, Harriman NY, February 1992. DARPA, Morgan Kaufmann.

[4] Julia Hirschberg and Christine Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting*, Santa Cruz, 1996. Association for Computational Linguistics.

[5] Julia Hirschberg and Christine Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998. ICSLP98.

[6] Julia Hirschberg and Christine Nakatani. Using machine learning to identify intonational segments. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Palo Alto, March 1998.

[7] Julia Hirschberg and Steve Whittaker. Studying search and archiving in a real audio database. In *Proceedings of the AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, Stanford, March 1997. AAAI.

[8] S.M. Katz. Estimation of probabilities from sparse data from the language model component of a speech recognizer. *IEEE Transactions of Acoustics, Speech and Signal Processing*, pages 400–401, 1987.

[9] Mehryar Mohri, Michael Riley, Don Hindle, Andrej Ljolje, and Fernando C. N. Pereira. Full expansion of context-dependent networks in large vocabulary speech recognition. In *Proceedings of ICASSP'98*. IEEE, 1998.

[10] Christine Nakatani, Steve Whittaker, and Julia Hirschberg. Now you hear it, now you don't: Empirical studies of audio browsing behavior. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998. ICSLP98.

[11] Fernando C. N. Pereira and Michael D. Riley. Speech recognition by composition of weighted finite automata. In Emmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*, pages 431–453. MIT Press, Cambridge, Massachusetts, 1997.

[12] John Pitrelli, Mary Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123–126, Yokohama, 1994. ICSLP.

[13] Gerard Salton, editor. *The SMART Retrieval System—Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[14] Kristie Seymore and Ronald Rosenfeld. Scalable backoff language models. In *Proceedings of the*

Fourth International Conference on Spoken Language Processing, volume 1, 1996.

[15] K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. ToBI: A standard scheme for labeling prosody. In *Proceedings of the Second International Conference on Spoken Language Processing*, pages 867–879, Banff, October 1992. ICSLP.

[16] Amit Singhal, John Choi, Donald Hindle, David Lewis, and Fernando Pereira. AT&T at TREC-7. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).

[17] E.M. Voorhees and D.K. Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999 (to appear).

[18] Steve Whittaker, John Choi, Julia Hirschberg, and Christine Nakatani. "what you see is almost what you hear: Design principles for accessing speech archives. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998. ICSLP98.

[19] Steve Whittaker, Julia Hirschberg, and Christine Nakatani. All talk and all action: strategies for managing voicemail messages. In *Human Factors in Computing Systems: CHI '98 Conference Proceedings*, Los Angeles, 1998.

[20] Steve Whittaker, Julia Hirschberg, and Christine Nakatani. Play it again: a study of the factors underlying speech browsing behavior. In *Proceedings of CHI '98*, Los Angeles, 1998. Human Factors in Computing Systems: CHI '98 Conference Proceedings.