Searching Recorded Speech Based on the Temporal Extent of Topic Labels

Douglas W. Oard* **and Anton Leuski**
University of Southern California Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina Del Rey, CA 90292-6695
oard@glue.umd.edu, leuski@isi.edu

Abstract

Recorded speech poses unusual challenges for the design of interactive end-user search systems. Automatic speech recognition is sufficiently accurate to support the automated components of interactive search systems in some applications. Recognizing useful recordings among those nominated by the system is difficult, however, because listening to audio is time consuming and because recognition errors and speech disfluencies make it difficult to mitigate this time factor by skimming automatic transcripts. Support for the browsing process based on supervised learning for automatic classification has shown promise, however, and a segment-then-label framework has emerged as the dominant paradigm for applying that technique to news broadcasts. This paper argues for a more general framework, which we call activation matrices, that provide a flexible representation for the mapping between labels and time. Three approaches to the generation of activation matrices could be generated are briefly described, with the main focus of the paper being the use of activation matrices to support search and selection in interactive systems.

Introduction

Recorded speech is a linear medium in which rapid skimming is hard to support. So although search based on speech recognition can be efficient, selection from a set of retrieved recordings, each of which might be several hours long, would be a time consuming process. Two approaches to this challenge have emerged: passage retrieval, and visualization. The key idea in passage retrieval is to divide speech recognition transcripts into thematically coherent (and relatively brief) segments, assign a score to each segment, and then present metadata that describes each high-scoring segment in the result list (Wechsler & Schaüble 1995). Visualization-based approaches, by contrast, retain the integrity of the original recording, but suggest points at which to begin the replay, typically using a timeline visualization. These suggestions can be based on thematically coherent segments that indicate both the onset and the temporal extent of potentially useful segments (as in the AT&T SCAN system (Whittaker *et al.* 1999)) or they can indicate only the onset (as in the HP Labs SpeechBot system (Thong *et al.* 2000)).

All three of the examples cited above are based on the presence of terms (words and/or phrases) that occur in the speech recognition transcript. In this paper, we explore alternative techniques in which automatic text classification is used to label periods of time in the recording, with a greater degree of abstraction than would be possible using just the actual words that were spoken. Classification-based approaches seem particularly well suited for use with linear media, since topic labels are easily skimmed. For example, Merlino and Maybury found that selection decisions could be made could be made more than twice as quickly and with comparable accuracy for news broadcasts when based on automatically assigned topic labels than when based on closed-caption text (precision increased 19%, although recall decreased by 11%) (Merlino & Maybury 1999). Segmentation and topic labeling have been the focus of the Topic Detection and Tracking evaluations, in which technology for improving access to audio and print news sources has been assessed annually since 1998 (Wayne 2000). In the TDT evaluations, segmentation and classification are modeled as separable problems, in which the system first seeks to detect segment boundaries created during the production process, and then seeks to label each segment with one or more topics. The BBN Oasis system incorporates this idea, displaying topic labels adjacent to the speech recognition transcript. This decomposition is appropriate for broadcast news, where segment combination is a natural part of the editorial process by which a news broadcast is created, but the utility of such a decomposition for naturally produced speech is less clear. In this paper, we propose an alternative approach in which the span of each label is determined individually.

Activation Matrices

We begin by defining an abstract representation for the span of each label. Figure 1 illustrates this structure, which we call an “activation matrix.” The rows represent a set of topic labels, for which some relationships between the labels may be known. In this paper we will restrict our attention to the case in which the relationships form one or more hierarchies,

*Permanent address: College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 such as might be commonly found in a thesaurus or an ontology. The columns represent units of time, with the granularity of the temporal representation left unspecified at this point as an implementation detail. Since two words per second would be relatively fast speech, it seems unlikely that sub-second temporal resolution would be needed in any application. So it might be helpful to think of each cell as representing one second of time. Each cell of an activation matrix contains a single real-valued number that represents the likelihood that the label represented by the row should be assigned at the time represented by the column. The precise nature of these likelihood values is also left unspecified as an implementation detail, but for the remainder of this paper we will choose to think of them as probabilities. In Figure 1, the degree of shading in a box is intended to represent the probability that the row's label should be assigned at the column's time, with the darkest shades representing certainty.

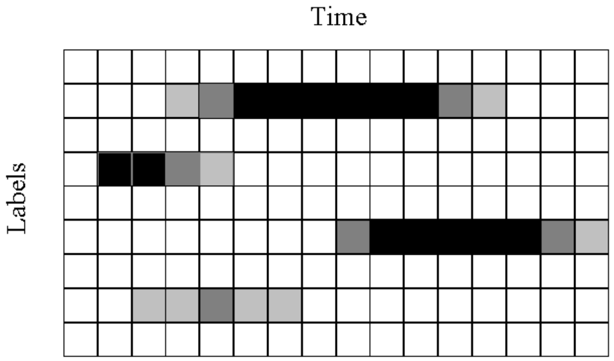


Figure 1: A sample activation matrix.

The first natural question to ask is where such activation matrices might come from. One possibility is that they might result from manual annotation of a collection by end users, as in the OntoLog system (Heggland 2002). Another way in which an activation matrix might be created is through a cascaded segment-then-label process using techniques demonstrated in the TDT evaluations. Figure 2 illustrates a third way in which such matrices might be created, relying on hand-labeled training data to train a model for the annotation process, and then applying that model to future data. Words contained in the speech recognition transcript are one example of a feature sequence that might be useful in such a classifier, but other features such as turn-taking behavior and silence may also prove valuable in some applications. Because our goal in this paper is to focus on how such activation matrices might be used, we will leave the detailed design of the model for future work (another of the "implementation details").

In most applications that we envision, the activation matrix would be quite sparse, with only a small fraction of the labels active at any particular time. Some form of compact representation would therefore be possible, and indeed would be necessary if we are to efficiently store any activation matrix with more than a few rows. Some variant of the inverted file index structure used in information retrieval would seem to be appropriate. For example, we could model the pattern of probabilities as a set of linear regions, recording the start time, initial value, and slope of each region. Each row could then be walked from left to right, easily reconstructing the value in any cell. Since it seems unlikely that appropriately smoothed probability values would exhibit high-frequency jitter in a real application, such a coding scheme would likely be both relatively compact and easily searched.

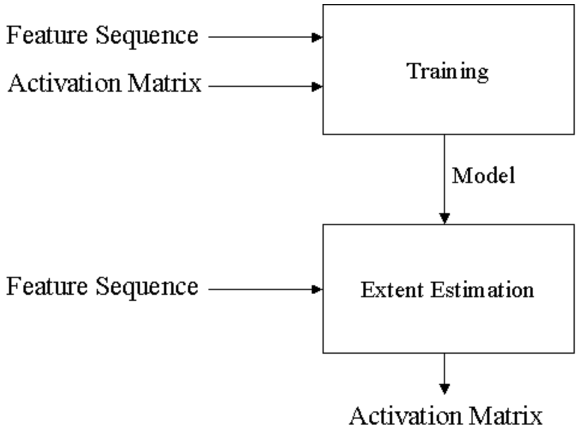


Figure 2: Supervised learning for activation matrices.

Finally, the question of what constitutes a good activation matrix is also important. For intrinsic evaluation, we could imagine a process in which some "gold standard" activation matrices are created by hand and then compared with those that are automatically generated. One possibility would be to use the L_1 norm of the matrix difference:

$$L_1 = \sum_{i,j} |M_{i,j} - A_{i,j}| \tag{1}$$

where $M_{i,j}$ is the probability at row i and column j of the manually created gold standard and $A_{i,j}$ is the corresponding value for an automatically generated activation matrix. The absolute value operator gives equal weight to errors in either direction; biased versions of this measure might also be used if, for example, precision were thought to be more important than recall in some application. It might also be be useful to normalize the measure in some way to facilitate interpretation. But the fundamental limitation of any intrinsic measure is that it can only reflect how well what is produced matches what is thought to be needed. To assess how well a an activation matrix actually supports some task, we must examine how it will be used. That is the focus of the next two sections.

Searching Activation Matrices

Modern text retrieval systems represent the contents of a document collection in a manner very similar to that of an activation matrix. In so-called "natural language" text retrieval, the rows typically represent terms (words or phrases), the columns typically represent documents or (for "passage retrieval") segments of documents, and the elements typically represent the degree to which a term describes a document (or segment). Mapping these concepts to an activation matrix is direct, with labels filling the role of terms, the segments possibly being shorter than is typical in text retrieval, and the degree of description being represented as a probability. We can therefore build on the substantial body of work on the design of text retrieval systems (Bae 1999).

There are two fundamental approaches to information retrieval: exact match retrieval and ranked retrieval. If the activation matrix contains only binary-valued elements (recording the presence or absence of a label at a time), then Boolean logic can be used in the query language to allow any possible combination of active labels to be specified. The natural result would be a set of contiguous spans in which the specified combination of labels is active. This set might further be ranked (e.g., in order of decreasing duration of the retrieved span) or clustered (e.g., with all spans from the same recording being shown on a single timeline).

Boolean logic offers an expressive query language, to which additional capabilities can be added using proximity operators and thesaurus-based query expansion. But Boolean logic has two key limitations that are important in practice. First, effectively expressing a query using Boolean logic requires a good deal of expertise (knowledge of Boolean logic, familiarity with appropriate iterative search strategies that minimize the all-or-nothing problem associated with overly general and overly specific queries, and a sufficient understanding of collection characteristics). An alternative approach is to use somewhat less expressive "natural language" queries to identify a broad range of potentially useful documents, emphasizing support for browsing by displaying the retrieved set in order of decreasing probability (or degree) of topical relevance to the query. The probabilities in the activation matrix offer a natural basis for creating such a ranked list.

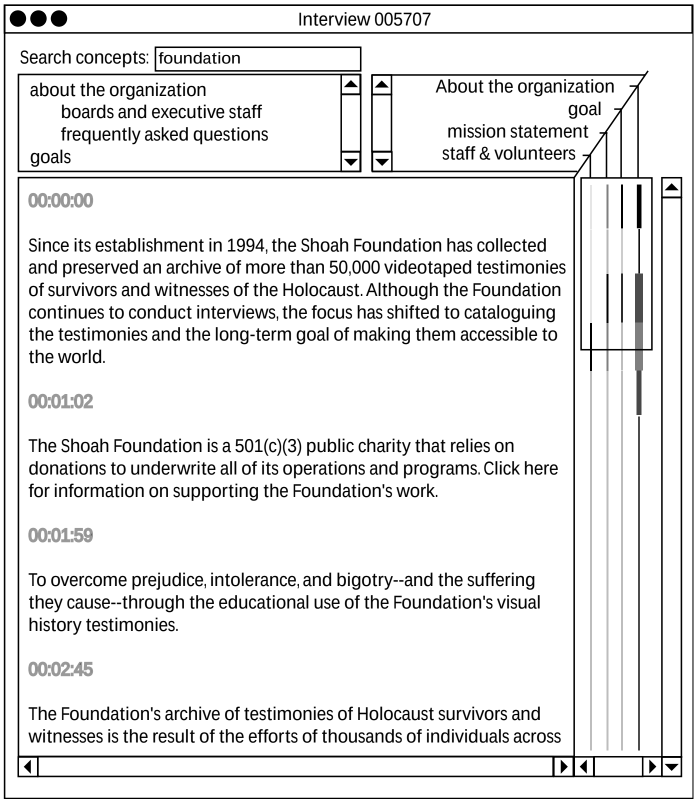
Ranked retrieval systems typically compute a score for each segment as follows:

$$s_j = \sum_i f(w_{i,j}, c_i) \qquad (2)$$

where s_j is the score assigned to segment j, $w_{i,j}$ is the degree to which label i describes segment j and c_i is the number of segments that are described by label i. The function $f(\cdot,\cdot)$ typically increases monotonically with w and decreases monotonically with c, capturing the intuition that relatively uncommon terms offer the most useful basis for ranked retrieval. Some systems combine ranked retrieval techniques with Boolean methods to provide expressive query languages that also include a "relevance ranking" capability. For example, modern Web search engines now often perform an implicit "AND" operation across the searcher's query terms and then ranked the returned results using some variant of (equation 2). The optimal balance between query complexity and interactive selection depends, of course, on how well we can support the process of browsing retrieved sets. That is the focus of the next section.

Visualizing Activation Matrices

We plan to visualize the activation matrix by showing what labels are associated with every temporal segment of an interview. Our assumption is that such a visualization would help a user to navigate, scan, and understand the content of the media stream. The total number of labels that are included into the activation matrix is very large – i.e., in the order of tens of thousand of labels. It is both impossible and impractical to present all the labels to the user at the same time. Thus the task of the interface breaks into two subtasks: (1) select an interesting subset of labels and (2) use this subset of labels to navigate the media stream. We prototype our interface to perform both tasks.



Search concepts: foundation

00:00:00

Since its establishment in 1994, the Shoah Foundation has collected and preserved an archive of more than 50,000 videotaped testimonies of survivors and witnesses of the Holocaust. Although the Foundation continues to conduct interviews, the focus has shifted to cataloguing the testimonies and the long-term goal of making them accessible to the world.

00:01:02

The Shoah Foundation is a 501(c)(3) public charity that relies on donations to underwrite all of its operations and programs. Click here for information on supporting the Foundation's work.

00:01:59

To overcome prejudice, intolerance, and bigotry--and the suffering they cause--through the educational use of the Foundation's visual history testimonies.

00:02:45

The Foundation's archive of testimonies of Holocaust survivors and witnesses is the result of the efforts of thousands of individuals across

Figure 3: A prototype of the interface.

Figure 3 shows a visual markup of the interface for an annotated document. The main part of the window is taken by a scrolling view with the audio transcript. The temporal segments are visible as individual paragraphs marked with the corresponding timestamps. The top right region of the window shows a list of four labels: "about the organization", "goal", "mission statement", and "staff & volunteers". These are the interesting labels selected to visualize the segment of the interview. We discuss how we use the labels for the visualization and then we describe how we select which labels to visualize.

Each label has a line associated with it that starts to the right of the label and then turns down following the transcript along the right edge. This is a timeline of the labeled concept activation during the whole span of the interview. The thin rectangle in the top right corner that surrounds about one third of each timeline corresponds to the visible portion of the transcript. We use the intensity of the gray color for the timeline to encode the likelihood of each concept being active during the corresponding portion of the interview. Dark gray corresponds to the segments where the concept is very likely to be present and light shaded pieces of the timeline indicate a likely absence of the concept in that part of the stream. For example, the concept labeled "staff & volunteers" is very likely to be present in the fourth time segment (i.e., the fourth paragraph) as the corresponding line segment is black and the same concept is very likely to be absent from the second time interval because the line is almost white.

Recall that the set of annotation labels has a hierarchical structure. In our example the label "about the organization" is the parent for the other three entries. The corresponding timeline (i.e., the rightmost timeline on the picture) is thicker. Both the timeline thickness and color for this parent concept are proportional to the number of the child concepts and the likelihood of the child concept being active at a given time interval. For example, the first segment is about the organization's goal and therefore it is about the organization. Both the child level concept ("goal") and the parent level concept ("about the organization") are active. The corresponding line segment is dark. During the second time segment only the parent level concept is active and the line is thin and dark. For the third segment both the "goal" and "mission statement" concepts are active, however the system is not very confident in its estimate and the parent level concept timeline is thick but more light-shaded.

Thus with a help from the visualization we can quickly establish that this interview describes an organization and it is more likely to go into details and describe the organization mission and members in the first half of the stream. We can also quickly zoom onto the part of the interview where the staff members are being described.

The list of labels at the top right corner of the window is a visible part of the list that contains all labels from the annotation hierarchy. The user can scroll the list using the scrollbar to the right of the labels or she scroll the same list with the small scrollbar at bottom right corner of the window. Scrolling the list brings other labels and their corresponding timelines into view.

The order of the labels in the list can be defined in several different ways. For example, we can order the labels by their relevance to the user's request or by their informativeness in describing the given interview or by the informativeness in describing the visible portion of the interview (so the list is reordered and visible labels change as the user scrolls the window). The user can also reorder the labels in the list by hand.

The top left area shows the system annotation data as a tree-like label hierarchy similar to file system browser. The user can search or browse the thesaurus for an interesting concept and then drag the label over to the right region. That reorders the list of visible labels and brings the selected labels into view.

The idea of mapping interesting areas of a document is well-known in information retrieval. For example, Hearst (Hearst 1995) divides documents into passages and visualizes individual passages as tiles. The tiles are shaded in proportion to the query term frequency in the corresponding passages. Byrd and Podorozhny (Byrd & Podorozhny 2000) color-code individual query terms and place the coded representations on the scrollbar directing the user to the most relevant parts of the document.

Our system prototype builds heavily on the ideas from the OntoLog system developed by Heggland (Heggland 2002). He explored the idea of helping a user to annotate temporal media with a hierarchical set of concepts. OntoLog visualizes the concepts as a tree structure similar to a file system browser. Each concept in the tree has a corresponding timeline attached to it. If a concept node is expanded, its timeline is shown as a thin line of uniform width. The timeline for a collapsed tree node has a variable thickness depending on the number of child concepts active at that time moment.

We expand the ideas expressed in the OntoLog system in three important directions:

1. OntoLog deals with human-assigned concepts – the concept either present or absent during a time segment. In contrast, our system works with automatically assigned concepts that have some probability to appear in the segment. These probabilities can assume any value between zero and one.

2. Heggland's system works with a small set of concepts (i.e., below eighty). We expect our interface to handle much large thesauri. We also believe that using a tree structure to order the concepts may pose a problem if the user tries to observe timelines for two concepts that are located in different parts of the tree – it might not be possible scroll the tree in such a way that both concepts are visible on the screen at the same time. Thus a mechanism for selecting and ordering a subset of the thesaurus for the visualization is very important.

3. In contrast to OntoLog, our system uses the concept timelines to describe and navigate the content of the stream transcript. It requires a tight link between the timeline visualization and the transcript. It also motivates the direction for the timelines – the natural flow of the text from top to bottom forces us to adopt the same direction for the timelines.

Conclusion

Anton will write this section.

Acknowledgments

The authors are grateful for Jon Heggland for helping us to see the link between ontologies and multimedia annotation, and to the MALACH team for their helpful comments on the ideas reported here. This material is based upon work supported by the National Science Foundation (NSF) under grant number 0122466. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1999. Modern information retrieval. New York: Addison Wesley.

Byrd, D., and Podorozhny, R. 2000. Adding boolean-quality control to best-match searching via an improved user interface. Technical Report IR-210, Department of Computer Science, University of Massachusetts, Amherst.

Hearst, M. A. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 56–66. ACM.

Heggland, J. 2002. OntoLog: Temporal annotation using ad-hoc ontologies and application profiles. In *Sixth European Conference on Research and Advanced Technology for Digital Libraries*.

Merlino, A., and Maybury, M. 1999. An emperical study of the optimal presentation of multimedia summaries of broadcast news. In Mani, I., and Maybury, M., eds., *Automated Text Summarization*.

Thong, J.-M. V.; Goddeau, D.; Litvinova, A.; Logan, B.; Moreno, P.; and Swain, M. 2000. SpeechBot:a speech recognition based audio indexing system for the web. In *Sixth RIAO Conference on Computer Assisted Information Retrieval*.

Wayne, C. L. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Second International Conference on Language Resources and Evaluation*, 1487–1494.

Wechsler, M., and Schaüble, P. 1995. Speech retrieval based on automatic indexing. In *Final Workshop on Multimedia Information Retrieval (MIRO '95)*.

Whittaker, S.; Hirschberg, J.; Choi, J.; Hindle, D.; Periera, F.; and Singhal, A. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In Marti Hearst, F. G., and Tong, R., eds., *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 26–33.