An Interactive Timeline for Speech Database Browsing

Benoit Favre, Jean-Franois Bonastre, Patrice Bellot

[1]LIA, University of Avignon, France

benoit.favre,jean-francois.bonastre,patrice.bellot@univ-avignon.fr

Abstract

Speech databases lack efficient interfaces to explore information along time. We introduce an interactive timeline that helps the user in browsing an audio stream on a large time scale and recontextualize targeted information. Time can be explored at different granularities using synchronized scales. We try to take advantage of automatic transcription to generate a conceptual structure of the database. The timeline is annotated with two elements to reflect the information distribution relevant to a user need. Information density is computed using an information retrieval model and displayed as a continuous shade on the timeline whereas anchorage points are expected to provide a stronger structure and to guide the user through his exploration. These points are generated using an extractive summarization algorithm. We present a prototype implementing the interactive timeline to browse broadcast news recordings.

Index Terms: audio browsing, timeline, information retrieval

1. Introduction

While recording speech as a digital media has brought the opportunity to build audio databases, there are few methods to access their content in convenient and efficient ways. Particularly, browsing speech databases is difficult for three main reasons. First, there is no natural description in the data that would help in retrieving material; an audio stream is basically a sequence of numbers quantifying the audio waveform. One may annotate the sound with metadata information such as topic, events, speakers, environment... This kind of annotation is expensive and cannot be performed at a low granularity. The second problem is a that audio streams cannot be skimmed as textual document by focusing on further paragraphs. It is more difficult to discriminate relevant and irrelevant spoken documents like skimming search engine results, because you cannot guess where relevant content is in a continuous stream and skip audio until this location. However, several solutions have been presented to search speech databases by content using textual transcripts of the discourse, generated by automatic speech recognition [1]. These search engines either provide full audio documents that are long to listen or small segments (sentences) that are affected by a loss of context.

In this paper, we present a complete automatic processing chain, form the signal to the user interface. Specifically, the user interface is designed for journalists and analysts to browse broadcast news continuously recorded during several years (section 3). The main contributions of this work are:

- An interactive playhead that involves multiple temporal scales to quickly locate a time point within decades of data.
- A timeline annotated with information density according to a user query.
- Anchorage points in this timeline to guide the user in his exploration of the database.

We also present a prototype implementation of this interactive timeline (section 4). It relies on automatic transcription and linguistic processing to take advantage of a conceptual representation.

2. Speech database browsing

Radio broadcasts represent a continuous and coherent source of spoken information. They are generally a good representation of main history events, strongly represented in the news headlines and less represented in other material such as variety programs. Archivists and analysts study past broadcasts looking for material related to an event or trying to take a global perspective on a current situation. Although broadcast news are annotated by hand-written metadata (named entities, topic words and small summaries or partial transcriptions), they lack long term analyses about groups of stories on a given topic. Those analyses could build a structure around the coverage of events and enable journalists to recontextualize information. For example, a political move may be a consequence of an unrelated event on which the politician was blamed not to react.

Journalists are interested in audio content for two main reasons. First, they have a moral duty to track information source and must identify the speaker, his exact words and the confidence on the material. Second, audio carries a lot more information than text about the environment, the speaker (state, emotion...) and the message. A manual transcription of every aspect of the audio signal is possible but no standard has been broadly adopted for this task. Moreover, listeners are used to getting those informations as a multi-modal input and interpreting them using a personal scheme that may not be efficiently replaced by textual representation.

User interfaces are needed to enable quick and efficient browsing of audio databases along different dimensions like time and topic. Those interfaces are expected to provide a continuum between a user need and the data, avoiding frustration by establishing a constant feedback on why the user's purpose could be misinterpreted by the system. We are interested in overcoming the search engine paradigm that enforces a segmentation of audio streams in documents and discards the relations between them like temporal order or topicality.

There are several ways to capture the user need that we classify as semantic, expressive and passive. First, a semantic approach allows the user to select conceptual elements (i.e. categories, keywords) from a closed list or using a text query. Then, an expressive approach lets the user select elements similar (or dissimilar) to what is sought (i.e. relevance feedback in information retrieval). Finally, in a passive approach, the system infers the user need by only observing his interactions. In this paper, we present a user interface that takes advantage of

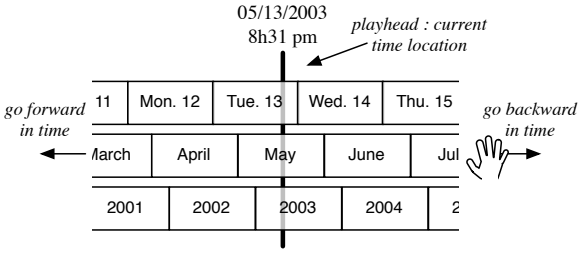

Figure 1: *The multiscale timeline used as a playhead. Locating a particular date is as easy as dragging one of the scale horizontally to place a temporal qualifier under the playhead at the center.*

the first two kinds of approaches. The system is designed to ease the implementation of the third one, but a suitable model remains to be deployed.

3. Interactive timeline

This section describes an interactive multiscale timeline driven by information retrieval and extractive summarization. The interface is built from a few constraints: (1) The user may want to quickly listen to audio at a given date in a database concerning decades. (2) The user may express an information need and the system will have to fulfill this need by providing appropriate feedback.

3.1. Multiscale playhead

In order to let the user locate time points on a big timespan, the timeline is represented at different granularities bound to natural qualifiers: years, months, days, hours and minutes. As shown in figure 1, each granularity is drawn as a slice of the timeline, graduated with corresponding labels. For example, the month slice contains a cycling list of month names (January, February, March...) separated by vertical lines. A central cursor represents the playhead at the current date. When audio is played, the timeline updates accordingly and the slices representing each granularity move in order for the cursor to reflect the updated timestamp. For instance, if the user wants to play audio one day earlier, he just drags the days slice to the right by one day (other slices are synchronized).

The multiscale timeline enables to locate time points using a natural representation of time for the user. Moreover, different dimensions can be represented on each slice and interpreted according to time at a specific granularity. This behaviour is interesting in the context of speech browsing as it brings the ability to show structural elements to guide the users (like folders structure a file system). In the context of broadcast news, the timeline could show event labels. Figure 2 illustrates an example of labels at the year and month levels for the war in Iraq.

For audio data, the annotation of events may be available to strongly structure the audio database. But we are mostly interested in the cases for which this kind of manual annotation is not possible. The next sections present automatic approaches to extracting and displaying structure information at the different granularities.

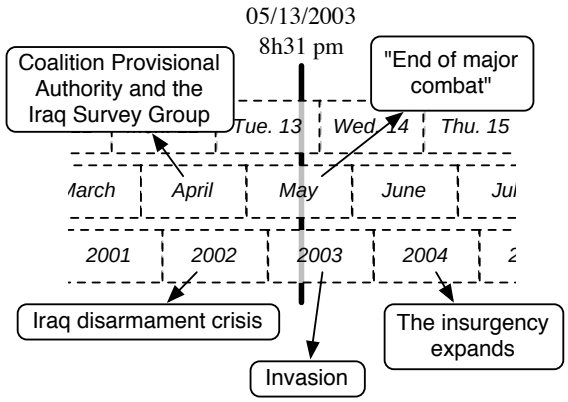

Figure 2: *Example of annotation of the timeline using labels from Wikipedia for the Iraq war (source:* `http://en.wikipedia.org/wiki/Iraq_War`*).*

3.2. Information density

We would like to annotate the timeline at different granularities with conceptual labels meaningful for the user. Extracting those kinds of high level labels directly from the audio is difficult because it requires a deep semantic analysis of the spoken discourse to detect and generalize events at the different granularities. This approach may lead to displaying labels irrelevant to the user. Heavy trends would emerge at the year level like well known or expected spans of history such as election, wars and disasters. The user is likely to be interested in events that may not have a yearly span: a specific air crash, a trial, a business event... In order to integrate this constraint, we let the user specify a text query to model his information need as in search engines. Then one could spot the query words in the audio stream and display them on the timeline. To go further, we fully transcribe the spoken content by automatic speech recognition. Thus, standard information retrieval models can estimate the relevance of source segments to the user information need. Then, the relevance may be used to annotate the timeline at each granularity. Using this representation, the user can locate topics interesting for him and can confront them according to time. This approach is less intuitive than labeling time-spans with conceptual items but it does not enforce wrong semantic decisions that would prevent the user to explore a given timespan. Representing relevance also has the advantage of straightforward granularity change at a low computational cost. Figure 3 illustrates the annotation of the timeline with "information density" according to a user need.

Even if the information density representation is more objective than conceptual labels, the user may be confused by the quantity of potential zones of interest in a timespan of several years. The user interface must be complemented by suggesting starting points to the user for his exploration.

3.3. Anchorage points

We define anchorage points as audio segments that contain meaningful material representative of the information that is relevant to the user need. If the user listens to audio at anchorage points first, he should spend less time exploring the database while looking for interesting segments. This approach acts in a similar way as a manual structuring in topics and sub-topics changing dynamically to reflect the user input query.

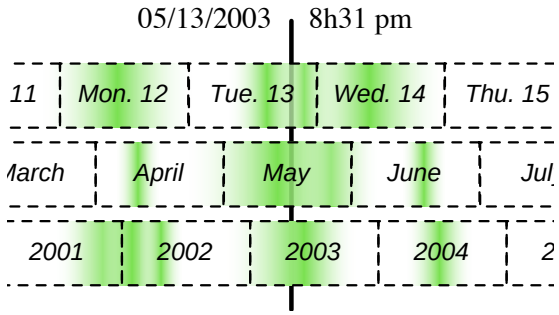

Figure 3: *Displaying the information density on the timeline according to the user need. Darker shades correspond to zones more relevant to the user need at each granularity.*

We follow a fully automatic approach based on query oriented extractive summarization to generate meaningful anchorage points. First, the speech stream is transcribed and cut into sentences. Then, using sentences relevant to the user query, the process builds a selection that maximizes the coverage of information while minimizing the redundancy. The temporal location of those sentences is displayed at each granularity: the user may click one of them to listen to the audio, or may just move the mouse over an anchorage point to highlight it on every scales.

Figure 4 shows the time locations selected as anchorage points to drive the user browsing experience. These elements bring a sharper structure to the relatively flat information density in order to keep the best of each approach: objectivity and meaning.

4. Implementation using the ESTER corpus

We implemented the method proposed in section 3 in a fully functional prototype using French broadcast news corpora (figure 5 shows a screenshot of the prototype). This prototype relies on basic techniques, but it will serve as a testbed for deployment and evaluation of more advanced approaches. The corpus is composed of 1700 hours of broadcast news from the ESTER evaluation campaign. The data[1] cover discontinuous periods from 1998 to 2004 with recordings from 10 minutes to 10 continuous hours. We decided to ignore the differences between continuous and discontinuous recordings: zones without data are considered irrelevant.

[1]The ESTER corpus contains overlapping segments from up to five sources. This problem of overlapping sources is not addressed yet in this work.

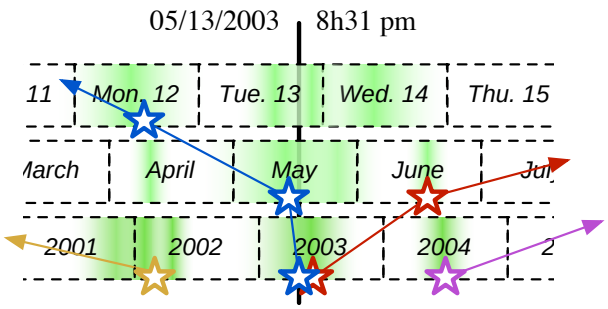

Figure 4: *Illustration of the anchorage points intended to guide the user in his exploration of relevant data. Lines link points represented at different granularities.*

The recordings have been automatically transcribed using Speeral [2] and Alize [3] toolkits developed at LIA for rich transcription of broadcast news. In addition to the textual transcription, the system provides a segmentation in macro-acoustic classes (speech, music...) and speaker diarization (identities and turns). The system was evaluated on 10 hours of broadcast news extracted from the corpus, transcribed manually. It performs at a word error rate (WER) of about 20%. The effect of WER can be limited, in an information retrieval context, using several techniques (N-best hypothesis indexing, phonetic matching and query expansion) as shown in the TREC [4] and TDT [5] evaluations.

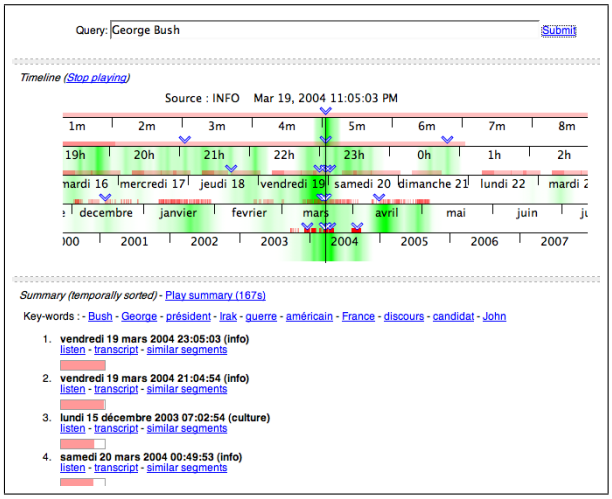

Figure 5: *Screen capture of the prototype: the user is interested in "George Bush" and explores data on March 19, 2004 at 11:05 PM. This date corresponds to the most relevant anchorage point detected. The prototype is available online.*

We indexed those transcripts at the sentence level to enable the system to locate relevant segments to a user query. We favor the sentences as indexing units as words alone are generally irrelevant to a user query (and far too numerous), and documents span on long durations and may tackle various topics including irrelevant parts (frustrating the user). The sentences are represented in a bag-of-words model including neighboring context as a simple linear interpolation (two sentences before and after the current sentence). Relevance to a user query is estimated using the Okapi model [6] which has a very good efficiency and good properties when dealing with variable sentence lengths. Other models like VSM [7] tend to favor short sentences with one or two words. Information density is computed as the normalized cumulative sentence relevance on a given timespan. It is represented on the timeline using a discretization on 1/20th of the unit for each granularity. Then density values need to be linearly interpolated to give the illusion of a continuous scale.

Anchorage points are selected within the n-best relevant sentences using modules from a text summarization system developed for the Document Understanding Conference (DUC) evaluation campaign [8]. The system is based on Maximal Marginal Relevance (MMR) [9] for the sentence selection. This greedy algorithm computes the similarity of a sentence to the user query (potential gain in coverage) and its similarity to

already selected sentences (potential increase in redundancy). Usually, information retrieval similarities (like Okapi or VSM) are used for the MMR similarity between sentences. Our system uses Latent Semantic Analysis (LSA) [10] to compute pseudo-semantic similarities. It consists in building a word co-occurrence matrix on an external corpora (here, 300M words from the French newspaper Le Monde) and reduce it to its main dimensions using Singular Value Decomposition (SVD). This process discards the less relevant dimensions and reveals some kind of word relations (it connects words that appear in the same context and words that have the same behaviour, like first names, for instance).

We added a few features to the prototype like the ability to get individual informations about the selected sentences (transcript, date, source...) and to listen to the "summary" formed by the anchoring points. The search engine and summarizer are server side processes while the user interface is a web page[2] driven by Javascript (the main layout), a Java applet (the timeline) and Flash (audio playback).

5. Discussion

We propose an innovative user interface for speech database browsing. This interface enables to explore the temporal distribution of information at different granularities. The timeline focuses on the order of events as spoken in the discourse, but a recording may deal with past or future events. These topics will be labeled with the date of the speech, whatever is their real date. Here, the timeline of discourse is opposed to the timeline of events, but these approaches are complementary and they may be represented in the same user interface (for instance: by highlighting differential zones between the two timelines). However, the timeline of events requires a deep analysis of discourse to extract and resolve time references.

In the presented interface, anchor points are selected by a summarization algorithm that discards time information on purpose. As time effectively plays a role in the distribution of information, we think that the user prefers anchor points decorrelated from time and that the timeline itself reintroduces the time information.

We have hypothesized that information density and anchor points would replace manually labeled concepts to structure the spoken information space according to a user need. But many users of our timeline claimed they would have found useful to be able to browse the database along topical dimensions. More than simple labels to time-spans, those dimensions need to be continuous and reflect the concept of specificity and generality to implement different granularity. Latent semantic models or ontologies like Wordnet (thanks to its hyperonymy and hyponymy relations) could provide strong bases in order to represent topical dimensions.

We are also interested in evaluating the proposed user interface. We already have conducted a small survey about how ergonomics may be improved and it appears that the interface confuses the users as every new interface, but after some experimentation, they find it natural and interesting. Whereas the individual modules of the interface were evaluated separately, we have not yet evaluated the relevance (in term of information) of the global user interface. It will require the definition of scenarios, like timing the operations needed by users to locate a piece of information or all segments about a given topic.

[2]The prototype is available at `http://www.lia.univ-avignon.fr/chercheurs/favre/timeline.html`.

6. Conclusion

In this paper, we have studied the problem of speech browsing in the context of broadcast news databases. We have presented an innovative user interface that enables one to quickly locate information on the timeline of discourse and recontextualize topics. The interactive timeline offers a multiscale view of information density according to the user need. While this density provides a smooth structure to the data, we add strong structure elements thanks to anchorage points that guide the exploration of zones of interest. We implemented information density using relevance scores from a search engine and anchorage point generation by an extractive summarization system.

We will extend this work by adding the opportunity to explore topical dimensions using semantic models and word relations. We will also try to include passive user need capture in the system, for instance as user profile recognition or using model adaptation to past sessions. Finally, an evaluation is needed to confirm the efficiency of the interactive timeline on real-world tasks.

7. References

[1] J.-M. V. Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain, "SpeechBot a Speech Recognition Based Audio Indexing System," in *Proc. of RIAO*, 2000.

[2] P. Nocéra, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, "The LIA's French Broadcast News Transcription System," in *Special Workshop in Maui (SWIM)*, 2004.

[3] D. Istrate, N. Scheffer, C. Fredouille, and J.-F. Bonastre, "Broadcast News Speaker Tracking for ESTER 2005 Campaign," in *Proc. of EUROSPEECH*, 2005, pp. 2445–2448.

[4] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Text REtrieval Conference (TREC)*, vol. 8, 1999, pp. 16–19.

[5] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer, 2002.

[6] K. Spärck-Jones, S. Walker, and S. Robertson, "A probabilistic model of information retrieval: development and comparative experiments," *Information Processing and Management: an International Journal*, vol. 36, no. 6, pp. 779–808, 2000.

[7] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic Query Expansion Using SMART: TREC 3," in *Text REtrieval Conference (TREC)*, 1994. [Online]. Available: citeseer.ist.psu.edu/37681.html

[8] B. Favre, F. Bechet, P. Bellot, F. Boudin, M. El-Beze, L. Gillard, G. Lapalme, and J.-M. Torres-Moreno, "The LIA-Thales summarization system at DUC-2006," in *Document Understanding Conference (DUC)*, 2006, pp. 131–138.

[9] J. Goldstein, V. Mittal, J. Carbonell, and J. Callan, "Creating and Evaluation Multi-Document Sentence Extract Summaries," in *Proc. of CIKM*, 2000.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.