VisPod: Content-Based Audio Visual Navigation

Qiyu Zhi, Suwen Lin, Shuai He, Ronald Metoyer, Nitesh V Chawla
University of Notre Dame
Notre Dame, USA
{qzhi, slin4, she1, rmetoyer, nchawla}@nd.edu

ABSTRACT

Current audio player interfaces generally provide brief information such as title and duration time and support basic playback control functions. These features alone are not sufficient for certain user tasks, such as quickly finding a previously-visited location or browsing the main topics covered in the audio content. We present *VisPod*, a visual audio player that visually displays the main topics and keywords extracted from the transcript. *VisPod* supports (1) audio content browsing, (2) topic-based and keyword-based navigation, (3) communication of transcript and speaker information in real time, and (4) content-based query. *VisPod* encodes audio as a donut chart comprised of topic segments, and uses text processing algorithms to segment the transcript into independent topics and utilizes a deep learning model to generate human-readable topic names. An informal study suggests users prefer *VisPod* over traditional audio playback approaches specifically with regards to its benefits for audio browsing and navigation.

ACM Classification Keywords

H.3.1 Information Storage and Retrieval: Content Analysis and IndexingLinguistic processing; H.5.2 Information Interface and Presentation (e.g., HCI): User InterfacesInteraction styles (e.g., commands, menus, forms, direct manipulation)

Author Keywords

Audio navigating; audio browsing; deep learning; topic generation; topic separation

INTRODUCTION

Audio is a ubiquitous type of multimedia for effectively conveying content such as interviews, lectures, and news. However, browsing and navigating such content are difficult with current audio players. For example, finding a previously-visited location is a difficult task as it is accomplished by trial-and-error by playing the file at various points until the desired content is found. The recent rapid development of Artificial Intelligence techniques such as Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) provides a content-based solution for audio exploration. In this paper, we present *VisPod*, a visual audio player that analyzes the audio transcript to generate a structure that allows users to easily browse and navigate the audio content. *VisPod* includes four key functions: **Audio content browsing**, **Topic-based and keyword-based navigation**, **Speaker identification**, and **Content-based query**.

IUI'18, March 13–16, 2018, Limassol, Cyprus

ISBN 123-4567-24-567/08/06…$15.00

DOI: http://dx.doi.org/10.475/123_4

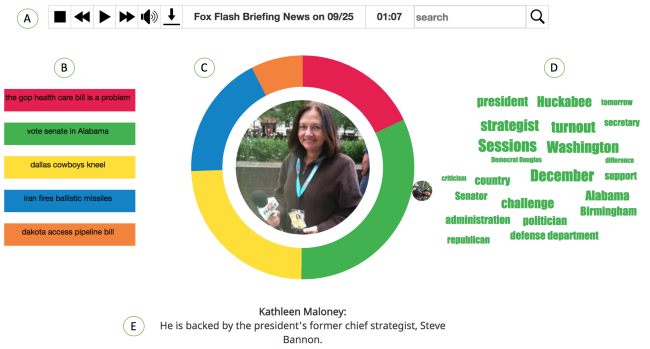

Figure 1. The figure shows the VisPod playing the Fox Flash Briefing News on 09/25.

VISPOD DESIGN

Figure 1 shows an overview of *VisPod* interface. The clock metaphor is a familiar representation that uses a spatial encoding for the quantitative time variable. We encode time duration as divisions of a donut chart to build on this clock metaphor and utilized a length encoding for topic durations. Time proceeds clockwise starting at the 12:00 position. We encode the current speaker with a small profile picture that moves clockwise around a donut chart indicating the time traversed in the audio file. Speaker information can also be viewed at the center of the donut chart. Color is an effective choice to encode nominal data. The same colors are used in both the "topic name" boxes, the donut chart segments, and all keywords for each topic. The initial word cloud on the right is generated from the complete transcript text with mixed colors representing all topics. The word cloud and its color will be updated for the current topic if the user clicks a "topic name" box or donut chart segment.

VisPod provides both topic-based and keyword-based navigation. A topic can be selected by either clicking on the topic title in the list or clicking on a topic segment in the donut chart. The audio will then start playing from this topic. Users can also click the keywords shown on the right to start playing

from the corresponding sentence location in this topic. In addition, time can be controlled by dragging the profile picture to the desired time location in the audio. As shown in Figure 1, users can browse the main topics on the left and keywords on the right. When a topic is clicked, the keywords cloud will be updated to show the content of the topic. Users can also browse the real-time transcript at bottom and speaker information at the center of the donut chart. *VisPod* also provides simple query function. Users can search the content in the audio, which will trigger a drop-down list including all suggested words. The audio will start to play the desired sentence after clicking a word.

IMPLEMENTATION

The implementation of *VisPod* consists of the following key procedures.

Audio and Transcript collection We manually transcribed the audio used for the *VisPod* demo. The required data format and *VisPod* demo can be accessed in supplemental materials. Once we obtain the transcript, we align it with the corresponding audio using P2FA [5] and CMU Sphinx Knowledge Base Tool [4].

Topic Segmentation *VisPod* provides the user with an overview of the audio's main topics. We use the TextTiling [1] algorithm to subdivide the transcript into the individual segments that represent subtopics. The result is a segmentation of the audio transcript into topics.

Topic Name Generation We utilize a RNN encoder-decoder model with LSTM units to generate the topic name of each topic segment. The dataset we use to train the model contains about 135,000 different news-headline pairs from 2016 to July 2017 and covers 15 news sources including the New York Times, CNN and the Guardian. Considering the computational complexity and model feasibility, we remove the data with more than 25 words in the headline and each news article is truncated into the first 50 words. All the text are then converted to lowercase and tokenized with the NLTK toolkit. An end-of-sequence symbol <eos> is appended to both ends of the headline and to the end of the news article. The numbers and dates in the data are transformed into symbols as "tag-num" and "tag-date". Finally, the data is randomly shuffled for the training process. As for our model, we first select the 40,000 most frequent words from our dataset as a vocabulary, then every word in the dataset is embedded into a distributed representation, where we use GloVe to initialize the embedding matrix. For the words not in the vocabulary, we first try to replace it with a word within the vocabulary if their cosine similarity exceeds a fixed threshold of 0.6. If there is no word in the vocabulary that is similar to the word, we mark it as a <*oov*> symbol. Second, the embedded text is fed into a 3-layer LSTM encoder and decoder. Each layer has 256 units and the dropout rate is set as 0.4 between the input and output gate of LSTM. Third, we implement an attention mechanism [2] to capture the long-range dependency in a long text, where 15% of the input is used to determine how much attention should be paid to the input and the remaining 85% accounts for the word prediction. The data is randomly divided into a set of 90% data for training and the other 10% for development. The training set is fed into our model for training with a batch size of 64 on a GPU machine. A smoothed BLEU is chosen to evaluate out model. As a result, the averaged smoothed BLEU score over the test data is 0.11, comparing to the BLEU scores 0.08 in the most related headline generation paper [2]. We also present three example results in Table 1.

Table 1. Example results for topic name generation model

Actual Topic Name	Generated Topic Name
north korea test fires two missiles, both fail	north korea fires ballistic missile
final oregon occupiers surrender to authorities, ending the refuge siege	oregon militiamen take refuge in oregon
white house officials suggest openness to immigration reform	white house : trump 's immigration plan would be replaced.

Keyword Extraction For each topic, *VisPod* presents all keywords as a word cloud to help a user better understand the content of the topic. After testing different methods described in prior work, we decided to use TextRank [3] for individual keyword extraction, which ranks the importance of a word using a graph-based ranking algorithm.

EVALUATION AND FUTURE WORK

We conducted an informal user evaluation to explore the usability of *VisPod*. We interviewed six participants who regularly listen to podcasts or other audio content. Overall, our users were able to quickly learn and get used to the navigation and browsing interactions. All the participants agreed that the clickable topic and keyword are intuitive and useful for navigating the audio. Participants also believed *VisPod* could be used in the real world and expressed their preference for a mobile-based *VisPod*. As one participant said: *lots of people listen to podcasts like NPR or other news, lots of them provide free transcripts online, I think you should work with them and offer the users a new audio interface.*

For future work, we plan to integrate ASR techniques into *VisPod* and evaluate *VisPod* with a formal study. Additionally, the generated topics lack details in some cases. We will focus on expanding training datasets and generalizing our model to more audio genres.

REFERENCES

1. Marti A Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23, 1 (1997), 33–64.
2. Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712* (2015).
3. Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text.. In *EMNLP*, Vol. 4. 404–411.
4. Alex Rudnicky. 2010. Sphinx knowledge base tool. (2010). `http://www.speech.cs.cmu.edu/tools/lmtool.html`
5. Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123, 5 (2008), 3878.