VisPod: A Visual Audio Player for Content Exploration

Qiyu Zhi*  Shuai He†  Ronald Metoyer‡
University of Notre Dame

Figure 1: The figure shows the VisPod playing an episode of a podcast named “The Model Health Show with Shawn Stevenson (TMHS)”: (A) shows the title and duration time. (B) lists all topic segments encoded by color. (C) displays the audio as a donut chart with each segment representing a topic corresponds to the topic lists at left. The small profile picture stands for the current speaker and its position indicates the current play time. Keywords of this topic are shown in the center. The importance of the keywords is encoded by word font size. (D) lists all speakers in the audio and the current speaker will be amplified and highlighted. (E) includes a play/pause button and the real time transcript.

Abstract

Current audio player interfaces generally provide brief information such as title and duration time and support basic control functions including play, pause, rewind, and forward. These features alone are not sufficient for certain user tasks, such as quickly finding a previously-visited location or browsing the main topics covered before listening. We present VisPod, a novel visual audio player interface to visually display main topics and keywords extracted from the uploaded transcript. VisPod supports overviewing the audio content in advance, navigating through topics, and browsing the content of each topic. An informal user study suggests that its features were perceived to be useful.

Index Terms: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Linguistic processing; H.5.2 [Information Interface and Presentation (e.g., HCI)]: User Interfaces—Interaction styles (e.g., commands, menus, forms, direct manipulation)

1 Introduction

Audio is a ubiquitous type of multimedia used for effectively conveying information including interview notes, lectures, and podcasts. The standard navigation functions in current audio players generally include play, pause, forward, and rewind. In addition, some podcast players such as iTunes also provide a title, duration time, and an introductory text. These features, however, are insufficient for tasks such as easily finding a previously-visited location or viewing the main topics covered before starting to listen. Specifically, revisiting a location in the conventional players is accomplished by trial-and-error by playing the file at various points until the desired content is found. Introductory paragraphs provided with the audio may be useful as a summary but too short to get an overall understanding of topics or too long to read in detail.

Researchers have addressed some of these problems with improvements to the basic audio player. For example, Uddin et al. use artificial landmarks to augment media players and allow users to revisit video locations effectively [7]. In addition, recent advances in speech recognition, such as Google Speech API 1, make other solutions feasible. For example, SpEX [1] uses a treemap to display the topical text structure and spoken words of lecture and presentation recordings. This allows students to identify and listen to the regions of audio that interest them while ignoring irrelevant regions.

We combine text processing algorithms and visual design to create VisPod (See Figure 1), a visual audio player to support four types of tasks: uploading an audio file (and its transcript), viewing the content summary, topic-based content exploration, and viewing the transcript in real-time. We use a donut chart to visually encode an audio file with each segment representing the main topic. To aid in understanding the content, we also display a word cloud with keywords corresponding to the current playing topic.

1https://cloud.google.com/speech/docs/
In the following sections, we describe the data processing techniques and the design of VisPod. We briefly discuss our preliminary study results before concluding and presenting directions for future work.

2 Data Processing

VisPod allows users to upload their own audio data and the corresponding transcript\(^2\). This section introduces the algorithms we use to process the transcript.

**Topic Segmentation:** VisPod provides the user with an overview of the audio’s main topics. We use the TextTiling [2] algorithm to subdivide the transcript into the individual segments that represent subtopics. TextTiling uses block comparison to compare the overall lexical similarity between adjacent sentence blocks. Through manipulating the block size, TextTiling is able to identify subtopic boundaries and determine the appropriate number of boundaries. The result is a segmentation of the audio into topics.

**Keyword Extraction:** For each topic, VisPod presents all keywords as a word cloud to help a user better understand the content of the topic. We explore both individual keyword extraction and keyphrase extraction. We use TF-IDF (term frequency-inverse document frequency) [3] for individual keyword extraction, which reflects how important a word is to a document in a collection or corpus. For keyphrase extraction, we use Rapid Automatic Key-Word Extraction (RAKE) [6]. RAKE splits the document text into an array of word phrases by the specified word delimiters. After calculating the importance score for each word, RAKE returns a list of key phrases sorted by aggregate importance score. We chose to use TF-IDF because individual keywords are more appropriate for presentation in the circular wordcloud when compared to the longer phrases from RAKE. The two most important keyword nouns are placed in the corresponding topic box (on the left) as a “topic title”.

3 VisPod Design

We use Munzner’s nested model [5] to guide the design decisions for VisPod. After interviewing several podcast listeners, we identified several domain tasks including overviewing the audio content in advance, navigating through topics, and browsing the content of each topic. A wide range of visual encoding and interaction decisions were considered and selected as described below.

3.1 Visual Encoding

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 utilize a length encoding for topic durations. Time proceeds clockwise starting at 12:00 position. We encode the current speaker with small profile picture that moves clockwise around a donut chart indicating the time traversed in the audio file. The complete circle of the donut equals the duration of the audio. Color is an effective choice to encode nominal data. We use divergent colors to encode different topics. Equivalent colors are used in both the “topic title” boxes and the donut chart segments for each topic. A word cloud is a simple and effective way to provide a fast and preliminary understanding of a body of text [4]. Rather than produce a word cloud for the entire audio, we show the word cloud for the current topic. In addition, the raw transcript will be updated at the bottom along with the audio. The current speaker’s name is highlighted on the right as the profile circle moves around the chart.

3.2 Interactions

VisPod is designed to allow users to quickly navigate through time and explore specific topics. Time is controlled by dragging the profile picture to the desired time location in the audio. 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. To facilitate smooth changes between topics and avoid change blindness, when a topic is clicked, the corresponding box will appear to translate to the central region of the donut chart, explode into the keywords-cloud, and then translate back. The audio will then start playing from this topic. Figure 2 shows that on clicking the topic contained in the orange-colored box or blue-colored box, the corresponding keywords-cloud is displayed and VisPod begins to play the audio starting from this topic.

4 Preliminary Study

An informal preliminary study with five graduate students provided qualitative feedback. They were all impressed with the innovative design and functional interactions. One of the participants remarked “seeing all keywords in a topic really helped me figure out which part of the audio I’m looking for”. They also thought VisPod would be more suited for use on mobile devices.

5 Conclusion and Future Work

In this paper, we presented a design for a visual audio player named VisPod. VisPod helps users understand the audio content at a glance and facilitates topic-based navigation through the audio content. The next step includes a formal user study to evaluate the usability and effectiveness of VisPod. Additionally, keywords are extracted based on word frequency and word-document co-occurrence. Integrating sound features into the keyword extraction process could be a promising future direction.

Acknowledgments

The authors wish to thank Yang Xu and other classmates in ND InfoVis class.

References


\(^{2}\)Required data format and VisPod demo can be accessed via: hci.nd.edu/v1spod/