July 2023

Gesture-driven Audio Bookmarks Powered by Large Language Model

Pol Henri Adrien Peiffer

Gesture-driven Audio Bookmarks Powered by Large Language Model

ABSTRACT

It is difficult to make a note of a fact, look up an entity, or perform other actions on audio content in the moment to enable remembering things for later. This disclosure describes techniques to create bookmarks for audio content such as podcasts, audiobooks, etc. with easy to perform gestures without interrupting the listening session. In response to the user performing the gesture, a capture flow is executed to transcribe, save, and make the audio content available for later search, reference, consumption, sharing, or browsing, e.g., via a bookmark. A large language model (LLM) can be utilized for various purposes such as to help the user search and revisit saved bookmarks via natural language queries to a conversational agent; to automatically generate titles for the bookmarked audio snippet; to summarize the audio snippet; to extract entities from the audio snippet; etc.

KEYWORDS

- Audio bookmark
- Audio summarization
- Audio search
- Large language model (LLM)
- LLM-based search
- Conversational agent
- Chatbot
- Natural language query
- Entity extraction

BACKGROUND

Audio from podcasts, audiobooks, social media, audio/video hosting websites, etc. contains rich information about places, people, facts, and ideas. The popularity of wireless earbuds (which give hands-free and eyes-free time to users), the widespread availability of

mobile data services, and the growth of streaming services have fueled the growth and availability of audio content.

However, it is difficult to act on audio content in the moment (while listening to audio content) and remember things for later. There are no easy ways to save a snippet of audio, or such features are specific to the app being used to access audio content. In many cases, there are no transcriptions of the content readily available. Audio content is generally not searchable and can therefore be difficult to find later, e.g., after a listening session. A related problem for hands-free/eyes-free content is that it can be hard for a user to act on information in the moment (e.g., remember a fact, look up a product, person, or topic) without interrupting the listening session.

DESCRIPTION

This disclosure describes techniques to create bookmarks that point to portions of audio content (e.g., podcasts, audiobooks, etc.) that the user has just heard. Upon executing a predetermined gesture, e.g., double-tapping an earphone, touching a button on a screen, etc. a capture flow is executed to transcribe, save, and make the audio content available for later activities such as search, reference, consumption, sharing, or browsing. The audio bookmarking techniques described herein enable users to save a particular timestamp of audio content they are listening to along with a note that can be saved in note taking or recording apps.

A large language model (LLM) can be leveraged to accelerate the searching and revisiting of the saved bookmarks using natural language querying. The LLM can also be used to automatically generate titles for bookmarked audio snippets, summarize audio snippets, extract entities from audio snippets, etc. Named entities extracted from an audio snippet can later be used for searching and for user interface building. For example, an extracted entity can inform a people/product search filter that can directly link to search results, panels, shopping, etc.

Information on entities can be automatically pulled in and displayed. For example, for people entities, their biographies can be automatically pulled in and placed on hover in the user interface. For product entities, their prices and reviews can be on hover. For place entities, such as performance venues, their description, reviews, and ticket prices can be on hover.

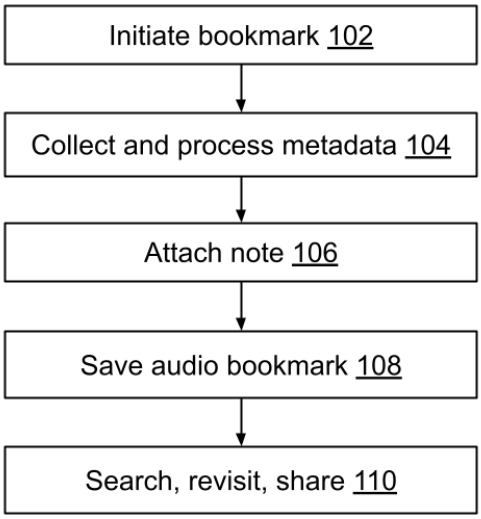

Fig. 1: User flow

Fig. 1 illustrates an overview of user flow with audio content and bookmarks, per techniques of this disclosure. Upon listening to an interesting section of audio content that they want to remember, a user initiates bookmarking (102) by performing a gesture or by other input method. In response to the user input, content metadata and timestamps are collected and processed (104). The user can optionally add a note (106) and save the note and the audio bookmark (108). The note and the audio bookmark can be saved to a user-selected destination, e.g., apps for note taking, recording, documentation, etc. The user can search, revisit, share, or browse their audio bookmarks (110) at any later time. Various components of the user flow are described in greater detail below.

Initiate bookmark (102): Features are provided to enable the user to initiate a bookmark in various ways, e.g., via a user interface button on a screen; via an earbud gesture (e.g., tapping a certain number of times on the earbud); via a voice command ("bookmark this!") that leverages existing hotword and virtual assistant capabilities; etc. A bookmark can thus be added without interrupting the user flow while listening to the audio.

Collect and process metadata (104): Examples of metadata that can be collected include the title of content that is playing; the artist or author of the podcast or audio content; the timestamp of the bookmarked content; a deep link to the app and to the specific content at the timestamp; etc. To account for latency between the start of interesting content and the user's decision to bookmark the content, the timestamp can be grabbed at a certain interval (e.g., 10-30 seconds) before the instant of bookmarking.

Metadata processing can be done using machine learning techniques (e.g., a large language model), and can include:

- *Title generation*: a concise title for the snippet is created based on the transcript and can be used as the metadata title, the snippet (bookmarked audio) title, or as a subtitle of the larger podcast or audiobook.
- *Summarization*: a summary or TLDR (too long; didn't read) of the snippet is created using the transcribed text and additional metadata such as title, author, etc.
- *Entity extraction*: named entities are extracted from the snippet for later use in searching and user interface building. For example, an extracted entity can inform a people/product search filter that can directly link to search results, panels, shopping, etc. Information on entities can be automatically pulled in and displayed. For example, for people entities, biographies can be automatically pulled in and placed on hover in the user interface. For

product entities, prices and reviews can be on hover. For place entities, e.g., performance venues, their description, reviews, and ticket prices can be on hover.

The above-described collection and processing of metadata is enabled by ephemeral transcription, e.g., a continuous audio buffer of a certain length, e.g., 30 seconds, is stored in a private, secure enclave, with permission from the user. Upon the user initiating a bookmark, the audio buffer is transcribed.

Attach note (106): The user can attach a note via voice (e.g., in a hands-free, screen-off manner) or via touch (e.g., with the screen on), each of which is explained in greater detail below.

- Attaching a note using voice: An example interaction with the virtual assistant is illustrated in Fig. 2.

Preamble Dialogue

```
User: Bookmark this.
Virtual assistant: Creating audio bookmark.
VA: Would you like to add a note?
```

At this point, the dialogue can branch, with example branches illustrated below.

Branch 1

```
U: Yes.
U: Really interesting book suggestion, follow up on that
VA: You saved the note: ‘Really interesting book
suggestion, follow up on that.’ Would you like to save it
or take a new note?
U: Save it.
VA: Bookmark saved.
```

Branch 2

```
U: No.
VA: Bookmark saved without note.
```

Fig. 2: Attaching a note using voice

- Attaching a note via touchscreen can be done with a layover, as illustrated in Fig. 3.

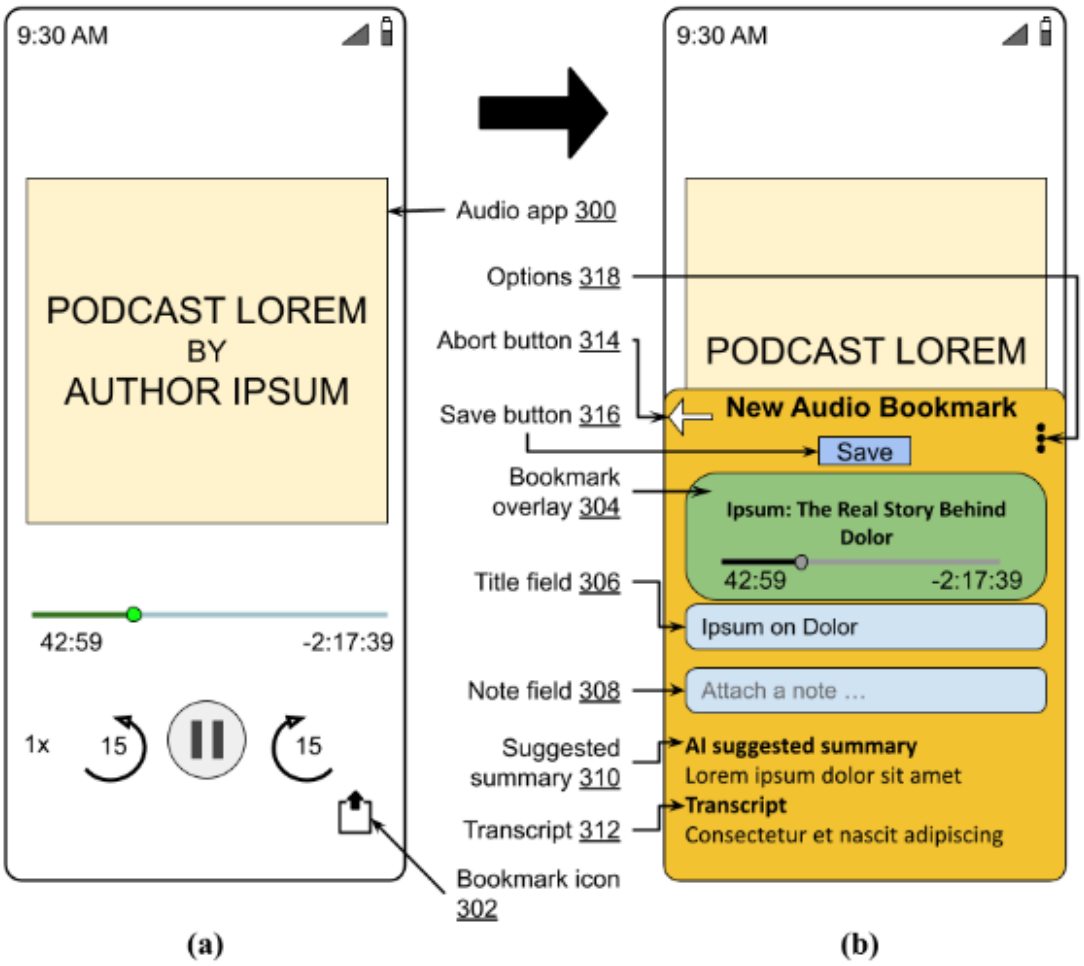

Fig. 3: Attaching a note via touchscreen: (a) Podcast user interface; (b) Audio bookmark overlay to enable the saving of the audio bookmark and associated notes

As illustrated in Fig. 3, a bookmark icon (302, Fig. 3a) can be provided on an app with audio content (e.g., audio app, 300) such that when tapped upon, a bookmark overlay screen (304, Fig, 3b) is displayed. The overlay shows timestamp and content information (e.g., thumbnail, author, etc.) as well as includes boxes for the user to enter a title (306), notes (308), etc. The title can be automatically generated and offered to the user for selection or editing. A transcript (312) of the last few seconds can be displayed alongside a generated summary (310), both of which the user can copy to fill in notes, title, etc. Additional user interface elements can

include a button to abort (314), a button to save (316), options (318) to indicate locations or apps where the bookmark and notes get saved, etc.

Save audio bookmark (108): The audio bookmark can be saved in a suitable user-selected location, e.g., a note taking app, a recording app, a documentation app, etc. If the bookmarked audio is saved to a note taking app, the bookmark can include data from the location and timestamp of the bookmarked audio in textual form, e.g., title, date, author, user-generated (and AI-assisted as necessary) note, transcript, summary, etc., alongside a link to the location and timestamp of the bookmarked audio. If the bookmarked audio is saved to a recording app, the bookmark can include an audio clip taken from the location and timestamp of the audio bookmark, alongside metadata such as title, author, etc. Linking can be done via a public feed such as via Really Simple Syndication (RSS).

Search, revisit, or share the bookmarked audio (110): The saved audio bookmark can be searched using standard search techniques as well as LLM-based search. For example, audio that is transcribed to text can be fed to an LLM for natural language querying and search, summarization, entity extraction, or other information processing tasks. A sharing experience custom to podcasts can be developed using RSS feeds. The user can thus easily organize audio information and recall or revisit it in the future.

In this manner, the techniques described herein enable users to structure and organize audio information without interrupting the user flow as they listen to audio content. Machine learning techniques are leveraged to automatically transcribe audio and organize audio information through traditional text-based search. A large language model (LLM) (e.g., via a conversational agent/chatbot interface) is leveraged to accept natural language commands for

various tasks related to the audio content such as to summarize, search, explain a concept in easier-to-understand terms, translate, etc.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's activity while listening to audio, audio bookmarks, audio snippets, natural language queries and other interaction with a conversational agent/chatbot, social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to create bookmarks for audio content such as podcasts, audiobooks, etc. with easy to perform gestures without interrupting the listening session. In response to the user performing the gesture, a capture flow is executed to transcribe, save, and make the audio content available for later search, reference, consumption, sharing, or browsing, e.g., via a bookmark. A large language model (LLM) can be utilized for various purposes such as to help the user search and revisit saved bookmarks via natural language

queries to a conversational agent; to automatically generate titles for the bookmarked audio snippet; to summarize the audio snippet; to extract entities from the audio snippet; etc.

REFERENCES

[1] "Unlock the knowledge in podcasts," https://www.snipd.com/ accessed June 24, 2023.

[2] Saigal, Rahul. "Three Android Apps to Help You Take Notes While Listening to a Podcast," available online at https://www.makeuseof.com/how-to-take-notes-while-listening-to-a-podcast-android/ accessed June 24, 2023.

[3] "Why Snipd is the best podcast app for saving and sharing highlights" https://www.youtube.com/watch?v=eXxKIkufPks accessed June 24, 2023.

[4] "How to take notes from podcasts with Snipd," available online at https://blog.snipd.com/how-to-take-notes-from-podcasts-with-snipd-6dd564d1c4ae accessed June 24, 2023.

[5] "Tatori — audio bookmark editor" available online at https://www.data.ai/en/apps/ios/app/tatori/?consentUpdate=updated accessed on Jun. 24, 2023.