March 2020

Just-in-time Delivery of Text-to-speech Playback

Jared M. Zimmerman

Joe Ashear

Just-in-time Delivery of Text-to-speech Playback

ABSTRACT

In an interaction between a virtual assistant and a user, it is often the case that the responses of the virtual assistant arrive just after the passage of critical events in the user's timeline. For example, navigational instructions to exit a freeway sometimes arrive just a bit too late to take the off-ramp. Some of this latency is attributable to the very nature of spoken audio playback: spoken audio playback being a linear medium, the length of time required for the spoken content to play can affect the accuracy of time-sensitive actions. This disclosure describes techniques that take into account the length of a response to be provided to the user such that the response is delivered in a timely manner with sufficient allowance for reaction time.

KEYWORDS

- Synthesized speech
- Text-to-speech (TTS)
- Virtual assistant
- TTS playback
- Time-sensitive response

BACKGROUND

In an interaction between a virtual assistant and a user, it is often the case that the responses of the virtual assistant arrive just after the passage of critical events in the user's timeline. For example, navigational instructions to exit a freeway may arrive just a bit too late to take the off-ramp. As another example, an assistant aiding a user cooking food may issue instructions ("boil the pasta for five more seconds and then take it out") that take so long to verbalize that the accuracy of executing the cooking step is affected. Similar examples are found in do-it-yourself (DIY) instructions, setting timers, timed cooking, games, etc. In general, any application where a virtual assistant issues real-time instructions or time-critical guidance can be subject to the problem of delayed delivery due to the length of the spoken instructions.

Some of the latency in delivering just-in-time responses is attributable to the very nature of spoken audio playback: spoken audio playback being a linear medium, the length of time for the spoken content to play can affect the accuracy of time-sensitive actions. In addition to playback time, the time taken to trigger and to queue the text-to-speech response affects the accuracy of time-sensitive actions.

DESCRIPTION

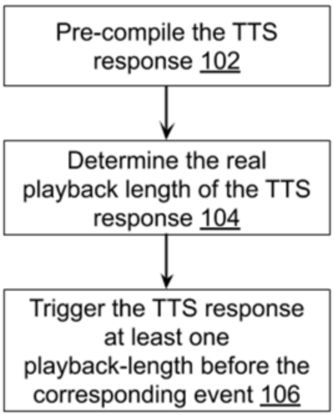
Pre-compile the TTS response 102
Determine the real playback length of the TTS response 104
Trigger the TTS response at least one playback-length before the corresponding event 106

Fig. 1: Just-in-time delivery of text-to-speech playback

Fig. 1 illustrates just-in-time delivery of text-to-speech (TTS) playback (synthesized speech, virtual assistant instruction, virtual assistant response, TTS response, etc.), per the techniques of this disclosure. The process of determining the right text string, e.g., virtual assistant response, and passing the string through an audio generation pipeline is referred to as compiling the synthesized audio. Given a text string, synthesized audio can be generated by combining phonemes from an audio library. Alternatively, a machine learning model can be trained to generate synthesized audio for the given text string. To minimize latencies arising out of client-server communication, the TTS response is pre-compiled (102) and stored on the client device. The playback length of the TTS response when played as audio is determined (104). The TTS response is triggered (106) at least one playback-length before the corresponding timed event. In this manner, the real and perceived accuracy of TTS playback on time-sensitive events and user flows is improved.

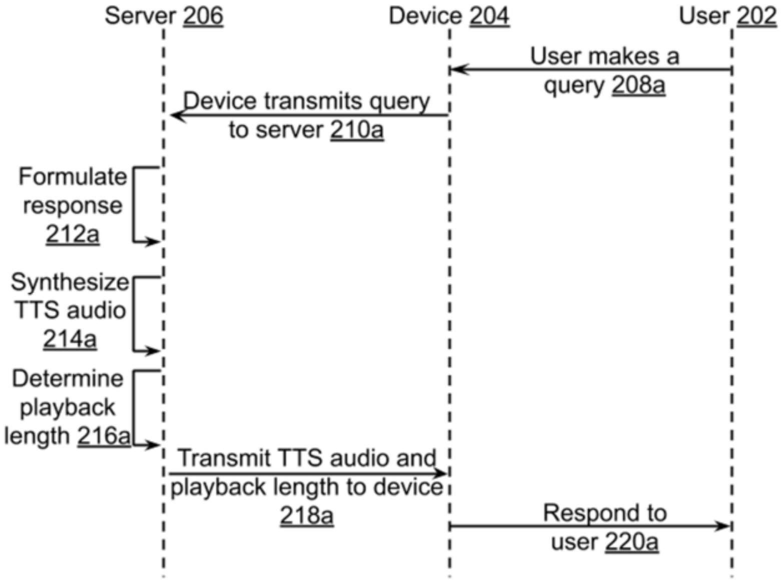
Server 206
Device 204
User 202
User makes a query 208a
Device transmits query to server 210a
Formulate response 212a
Synthesize TTS audio 214a
Determine playback length 216a
Transmit TTS audio and playback length to device 218a
Respond to user 220a

Fig. 2A

The tasks involved in generating a TTS response and determining playback length can be divided between a user device and a server in various ways. Fig. 2A illustrates an example where the tasks are largely performed at the server (206). A user (202) makes a query (208a) to a device (204), which transmits it to a server (210a). The server compiles an audio response, e.g., it formulates a text response (212a), synthesizes the TTS audio (214a), and determines playback length (216a). The server transmits the TTS audio and playback length (e.g., as metadata) to the device (218a). The device responds to the user (220a).

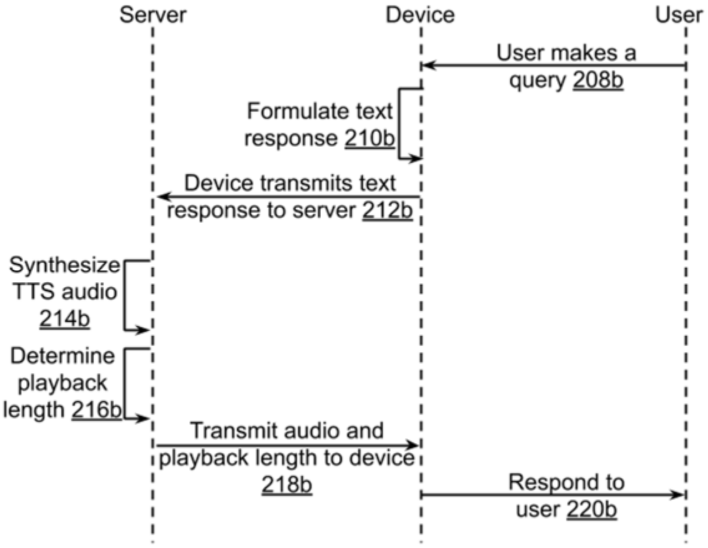
Server
Device
User
User makes a query 208b
Formulate text response 210b
Device transmits text response to server 212b
Synthesize TTS audio 214b
Determine playback length 216b
Transmit audio and playback length to device 218b
Respond to user 220b

Fig. 2B

Fig. 2B illustrates an example where the tasks are split between device and server. Upon receiving a user query (208b), a text response is formulated at the device (210b). The device transmits the text response to the server (212b), which synthesizes the TTS audio (214b) and determines playback length (216b). The server transmits the TTS audio and playback length (e.g., as metadata) to the device (218b). The device responds to the user (220b).

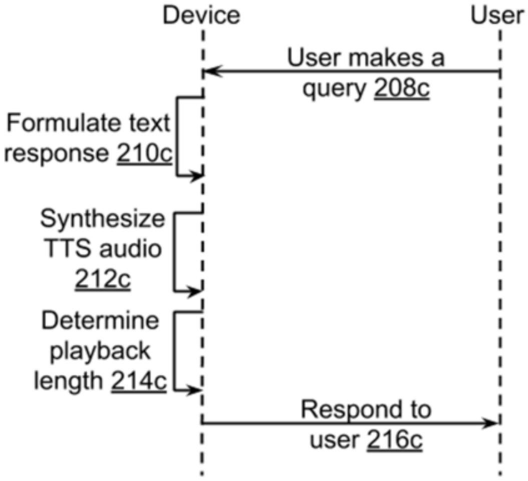
Device
User
User makes a query 208c
Formulate text response 210c
Synthesize TTS audio 212c
Determine playback length 214c
Respond to user 216c

Fig. 2C

Fig. 2C illustrates an example where the tasks are performed entirely on the device. Upon receiving a user query (208c), the device compiles a response, e.g., it formulates a text response (210c), synthesizes TTS audio (212c), and determines playback length (214c). The device responds to the user (216c). For example, this mode in which tasks are performed on the device, is suitable for local actions, e.g., user commands such as “turn down the screen brightness.” As mentioned before, the device response can be accelerated by pre-compiling the response, e.g., by performing the text processing and TTS generation steps (210c-214c) ahead of time.

While Figs. 2A-2C illustrate three examples of division of tasks between a server and a client device, the tasks can be divided in any other technically feasible way. The division of tasks can be based on factors such as the type or content of the query, local device processing capabilities, network reliability, network latency, etc. Further, the user is provided with options to choose whether query processing happens locally or remotely on a server. For example, the user can specify that certain types of queries be responded to locally, and other types of queries can be processed on the server. Queries are sent to the remote server only when the user has provided permission for server processing.

CONCLUSION

This disclosure describes techniques that take into account the length of a response to be provided to the user such that the response is delivered in a timely manner with sufficient allowance for reaction time.