May 06, 2019

Automatic selection of audio compression for spoken commands

Jordan Werthman

Glen Shires

Automatic selection of audio compression for spoken commands

ABSTRACT

Voice commands are commonly used for interaction with virtual assistant applications provided via user devices such as smart speakers, appliances, smartphones, etc. When a user provides permission, some voice-enabled applications upload the user's speech data to a server using lossless compression to enable server-based recognition of the user command. The lossless nature of the transmission can take up significant network resources and receiving a response from the server can take a significant amount of time when the user has a slow network connection. This disclosure provides techniques that enable faster transmission for server-side processing of user speech data while retaining recognition quality. Allowing loss in the transmitted audio reduces the resources required for speech data transmission. To ensure that there is no loss of quality, the user's environment is evaluated with user permission, to determine whether lossy transmission is feasible for the particular user speech.

KEYWORDS

- voice command
- voice query
- audio compression
- digital assistant
- virtual assistant
- smart speaker
- smart display
- hotword
- speech recognition

BACKGROUND

Voice commands are commonly used for interaction with virtual assistant applications provided via user devices such as smart speakers, appliances, smartphones, computers, wearable devices, etc. When a user provides permission, some voice-enabled applications upload the user's speech data to a server using lossless compression to enable server-based recognition of the user command. The lossless nature of the transmission can take up significant network resources and receiving a response from the server can take a significant amount of time when the user has a slow network connection. The quality of speech recognition is also negatively impacted by a low signal-to-noise ratio in the user speech, e.g., the presence of high amounts of background noise from various sources.

Inaccurate speech recognition has a negative effect on user experience since the response from a virtual assistant for an incorrectly recognized voice command may not match user expectations. Further, this can also impact voice-based activation of devices, e.g., by the use of a hotwords that are used as a phrase to "wake up" a device, when the server-based processing fails to recognize that the user uttered the hotword.

DESCRIPTION

To provide high quality recognitions while reducing transmission time, this disclosure makes a determination of the suitability of lossy compression before transmitting the user speech for server processing. By evaluating the likelihood of accurate recognition with lossy compression, the techniques enable reduction in transmission time. If the likelihood is low, the user speech is transmitted using lossless techniques. Factors such as available network speed and the quality of user speech, as indicated by signal-to-noise ratio and/or distortions are utilized in determination of the suitability of lossy transmission.

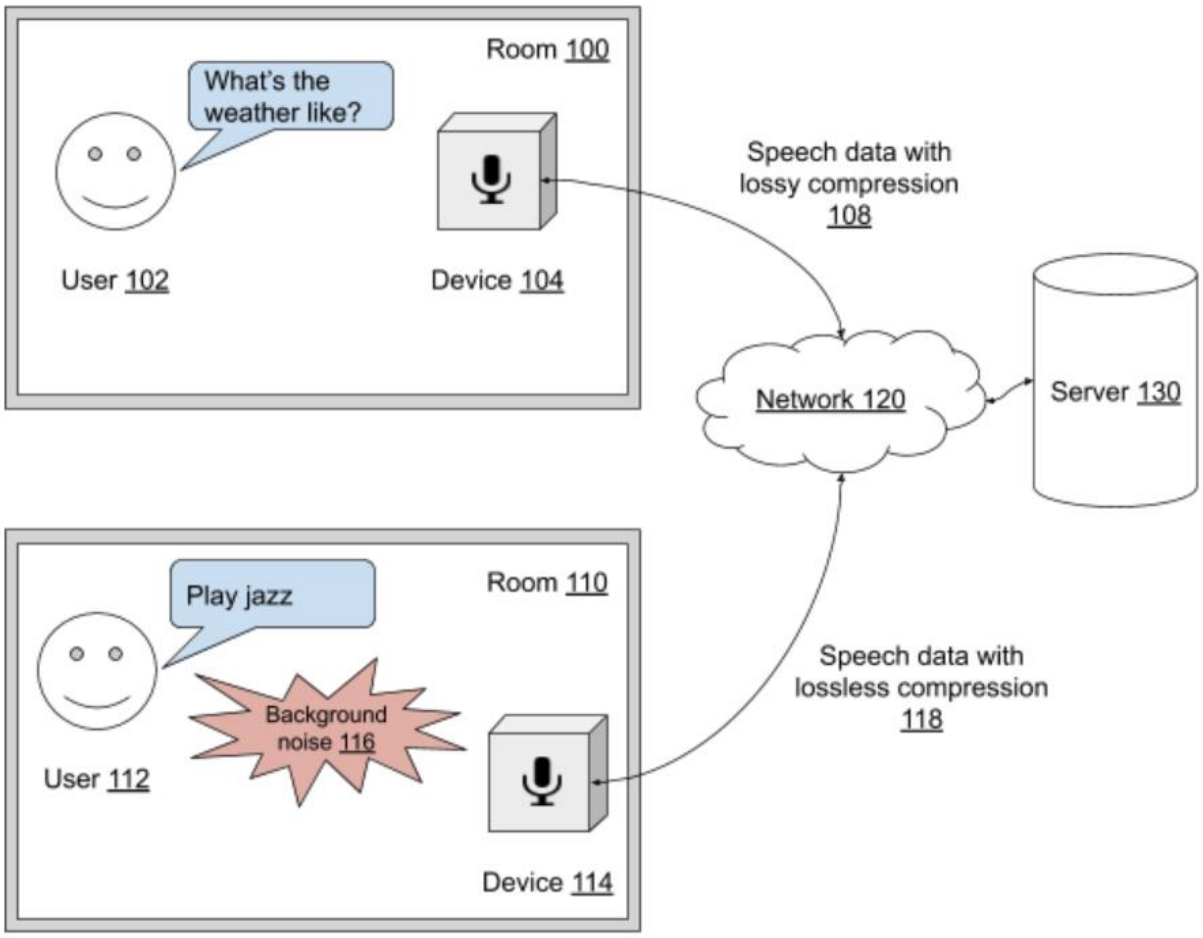

Fig. 1: Selection of compression mechanism for transmission of speech data

Fig. 1 illustrates example scenarios in which the techniques for selection of compression mechanism as described herein are utilized. As shown in the figure, a first user (102) is in a quiet room (100) and issues a voice command ("What's the weather like?") to a device (104). The device detects that the user is nearby and that there is little to no background noise. Thus, the device determines that lossy compression can be utilized for transmission speech data to server (130) via network (120). Lossy compression can be utilized selectively, e.g., when the network performance is low, when using the network resource has a high cost, etc. When the received speech is clear, e.g., has minimal distortions and/or high signal-to-noise ratio, and is likely nearfield (spoken from close to the device) as opposed to far field (which introduces

reverberation, echoes, and/or fragmentation), lossy compression is applied prior to transmission to the server. Owing to the high quality of the received speech, the compression has minimal impact on speech-to-text accuracy and word error rate (WER). Sending speech data using lossy compression saves network resources and can be completed in less transmission time, enabling a response to be provided faster.

A second user (112) is in a room (110) that includes background noise (116). Also, the user is far from a user device (114) that receives a spoken command, e.g., “Play jazz.” In this example, the speech is unclear, e.g., has low signal-to-noise ratio, has distortions such as echo produced in a farfield environment, etc. and it is determined that lossy compression can have a negative impact on recognition accuracy. In this case, the speech data is sent using lossless compression or high quality lossy compression.

Automatic selection of lossless vs. lossy compression based on various factors as described above allows for user queries to be completed faster, with reduced bandwidth usage, when feasible, while still delivering high quality results to the user. Use of the techniques reduce latency in general without negatively impacting accuracy.

Many voice-activated assistant applications and devices such as smart speakers, smart displays, etc. utilize a client-side hotword (wake-word) detector to detect when a user says a hotword that activates the assistant application. Many hotword detection techniques calculate a probability that the audio received from the user did indeed match the hotword. A higher probability is indicative that the speech is clear. The probability can be compared against a predetermined threshold to select whether to transmit the audio/data via a lossy compression (e.g., when probability is above the threshold) or a lossless or less-lossy compression (e.g., when probability is below the threshold).

Various other techniques can be utilized to determine various factors that are used to select lossless or lossy compression for transmission of speech data. Some of these techniques include:

- Use of audio processing techniques can be utilized to determine reverberation/echo characteristics in the received audio during the period that the hotword is detected. The determined characteristics can be used as a selection factor.
- Speech processing techniques can be applied locally on a client device to determine clarity of features in the audio during the period the hotword is detected, and the determined clarity level can be used as a selection factor.
- Speech processing techniques can also be used to determine the average background noise in the audio before the period the hotword is detected as compared to the period when the hotword is detected.

The evaluation can be performed separately or jointly on multiple channels of audio, e.g., audio captured from multiple microphones. Two or more of the above techniques can be combined. As an alternative to selecting between lossless and lossy compression, the described techniques can also be used to select between two or more levels of lossy compression. Multiple levels of lossy compression can be utilized, each with a corresponding threshold. For example, multiple thresholds can be to select between three or more levels of lossy compression, or between lossless and two or more levels of lossy compression. Still further, a dynamic threshold can be implemented that changes based on prior usage, if permitted by the user, such as the accuracy level of prior speech recognition attempts.

The dynamic threshold can also be determined based on an estimation of network speed, e.g., determined based on prior audio/data uploads via the network, a speed test that periodically

sends packets over the network, or by monitoring the local network to determine the volume of traffic from other devices.

The described selection techniques can also be utilized for other purposes, e.g., for server-side audio processing, conferencing between two or more clients, etc. The described techniques can be utilized in smart speakers, smart displays, televisions, soundbars, home appliances, in-car appliances, 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 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 provides techniques that enable faster transmission for server-side processing of user speech data while retaining recognition quality. Allowing loss in the transmitted audio reduces the resources required for speech data transmission. To ensure that there is no loss of quality, the user's environment is evaluated, with user permission, to determine whether lossy transmission is feasible for the particular user speech.