February 13, 2019

Dynamic adjustment of hotword detection threshold

Tanmay Wadhwa

Neil Dhillon

Dynamic adjustment of hotword detection threshold

ABSTRACT

Many devices allow users to speak a hotword to activate the device, e.g., a virtual assistant application, which then responds to the user command. With user permission, incoming speech data is analyzed to determine whether a hotword was uttered. First, coarse hotword detection is performed. If coarse detection indicates that the hotword was spoken, fine hotword detection is performed to confirm that the hotword was spoken. Per techniques described herein, when fine hotword detection is unsuccessful, the threshold for fine hotword detection is reduced for a short time window. Such reduction improves the likelihood of recognition of the next utterance of the hotword, and can reduce consecutive false negatives. Further, the response from the device is adjusted to improve user experience.

KEYWORDS

- virtual assistant
- smart assistant
- smart speaker
- hotword
- activation phrase
- threshold adjustment

BACKGROUND

Many devices allow users to speak a hotword or activation phrase to activate the device, e.g., a virtual assistant application, which then responds to the user command. The accuracy of hotword detection varies based on environmental factors. For example, hotword detection is less likely to be successful in a noisy environment, when multiple users speak simultaneously, etc.

Hotword detection can also fail due to differences in the user's voice, e.g., change in tone. In certain circumstances, e.g., when initial hotword detection fails, there is often a change in tone for the next utterance of the hotword, e.g., since the user more likely to be frustrated due to lack of response from the device.

DESCRIPTION

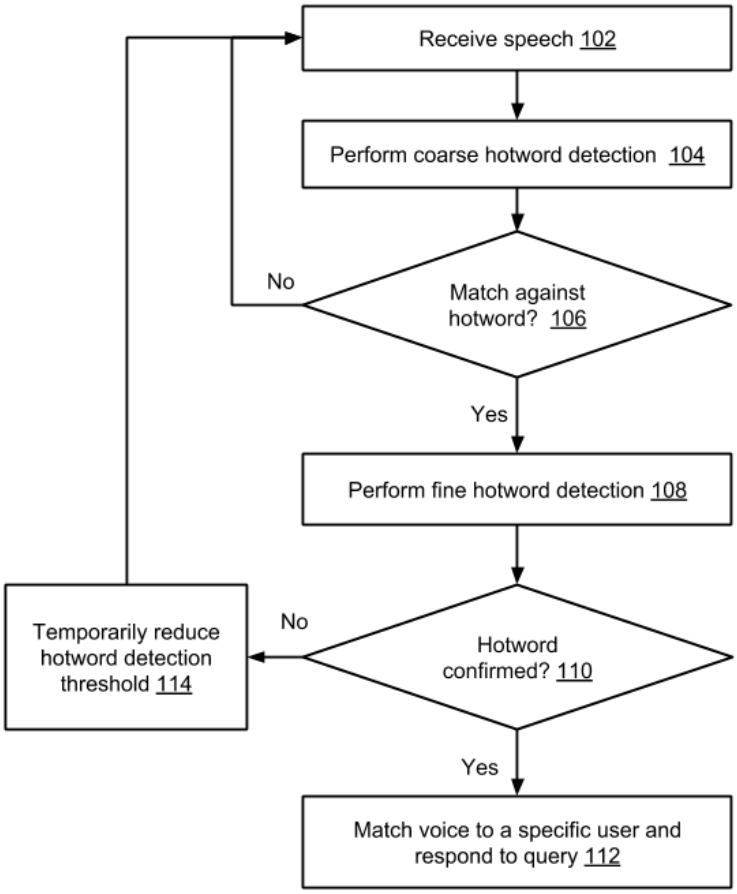

Fig. 1: Dynamic Threshold adjustment of hotword detection

Fig. 1 illustrates an example hotword detection process that can reduce the likelihood of false negatives. Hotword detection is performed with user permission to detect user speech and to automatically analyze the speech. The user is provided options to turn off speech detection and to disable voice queries entirely.

A device that is configured to respond to user-spoken hotwords, e.g., smartphone, tablet, smartwatch or other wearable device, smart appliance, etc. receives user speech (102). For example, when the device is configured with a virtual assistant, the speech is analyzed to detect whether it includes a preconfigured hotword for activation of the virtual assistant. For example, the hotword can be a predetermined phrase such as “Hey Helper.” Some devices can include a low-powered digital signal processing chip that can be utilized for hotword detection.

First, coarse hotword detection (104) is performed. Coarse hotword detection is a computationally inexpensive operation that compares received speech with the predetermined phrase (hotword). If the coarse detection indicates a match against hotword (106), fine hotword detection is performed.

Fine hotword detection (108) is computationally more expensive than coarse hotword detection. When users permit, fine hotword detection can be performed using a model customized to specific users (108), e.g., a machine-learning model that is trained on prior user speech data. Further, when user permit, fine hotword detection can include speaker identification. Fine hotword detection is performed using a threshold to match the received speech with the hotword. If fine hotword detection indicates a match (110) that meets the threshold, a response is provided (112) to the query in the user speech. When speaker identification is enabled, the response is based on matching the incoming speech against a specific user’s voice.

Per techniques of this disclosure, if fine hotword detection indicates that there is no match, the threshold is reduced temporarily (114), e.g., for the next 10 seconds. Such adjustment is based on the intuition that near-misses indicate a possibility that the user uttered the hotword that was not recognized, e.g., due to noisy environment, tonal variation, etc. The reduction in

threshold is temporary to prevent inadvertent triggering of the device at a later time, far removed from the initial utterance. If the user repeats the voice query within the temporary time period, the device is more likely to recognize the hotword owing to the adjustment in the threshold.

Such adjustment can prevent multiple consecutive false negatives where the device fails to activate in response to utterances of the hotword which improves the user experience. Further, when hotword recognition succeeds at a second (or later) attempt with the reduced threshold, the virtual assistant can be configured to adjust the response provided. For example, the virtual assistant can state “I’m sorry! didn’t quite get you earlier. How can I help?” if the hotword is not immediately followed by a query, or can provide an apology followed by an answer to the user’s query.

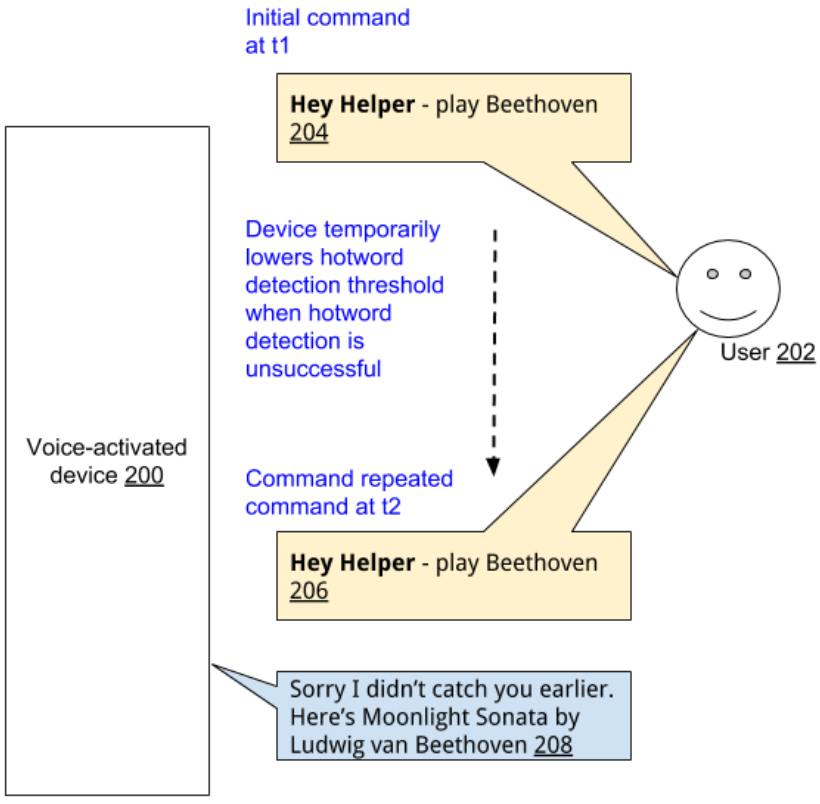

Fig. 2: Device with dynamic hotword detection

Fig. 2 illustrates an example operation of a voice-activated device (200) that implements a virtual assistant with hotword detection techniques as described herein. The device is configured to recognize the hotword “Hey Helper.” At time t1, the user utters a command that includes a hotword, e.g., “Hey Helper, play Beethoven” (202). The voice-activated device fails to recognize the initial hotword dictated by the user. However, as described above with reference to Fig. 1, if fine hotword detection is unsuccessful due to a relatively small mismatch, the hotword detection threshold is lowered temporarily.

When the user repeats the command at time t2, the hotword is recognized successfully (206) and the user command “play Beethoven” is recognized. The device responds to the user with an apology and provides a response to the user command.

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

The disclosure describes techniques to dynamically adjust threshold for hotword detection in devices with a virtual assistant. With user permission, incoming speech data is analyzed to determine whether a hotword was uttered. First, coarse hotword detection is performed. If coarse detection indicates that the hotword was spoken, fine hotword detection is performed to confirm that the hotword was spoken. Per techniques described herein, when fine hotword detection is unsuccessful, the threshold for fine hotword detection is reduced for a short time window. Such reduction improves the likelihood of recognition of the next utterance of the hotword, and can reduce consecutive false negatives. Further, the response from the device is adjusted to improve user experience.

REFERENCES

1. Moray, Neville. "Attention in dichotic listening: Affective cues and the influence of instructions." Quarterly journal of experimental psychology 11, no. 1 (1959): 56-60.
2. Nakane, Toshiki, Makoto Miyakoshi, Toshiharu Nakai, and Shinji Naganawa. "How the Non-attending Brain Hears Its Owner's Name." Cerebral cortex (New York, NY: 1991) 26, no. 10 (2016): 3889-3904.