July 2023

Low Latency and Energy-efficient Hotword Detection Based on Mouth Movement

D Shin

Low Latency and Energy-efficient Hotword Detection Based on Mouth Movement

ABSTRACT

Devices that provide virtual assistants that respond to spoken queries monitor the user's speech to detect a hotword to trigger the virtual assistant application. However, hotword detection requires some components of the device to be always on which make such devices power hungry. Always-on microphones are a problem especially for small devices such as hearables that have a small battery. This disclosure describes techniques that gate hotword detection with a lightweight neural network that estimates mouth motion signatures using signals from a low-power, always-on inertial measurement unit (IMU). Because the IMU has lower bandwidth than microphones, signals generated by the IMU can be processed by a neural network that is small enough to reside on the front-end processor of a low-powered device such as a hearable. Hotword gating can be done without waking up power hungry processors of the device. IMU-based gating can allow high-precision hotword detection to be achieved at very low power consumption.

KEYWORDS

- Mouth movement
- Lip movement
- Hotword
- Wake word
- Virtual assistant
- Hearable
- Smart headphones
- Earbuds
- Inertial measurement unit (IMU)
- Always-on microphone
- Low power computing
- Power budget

BACKGROUND

A common way to activate a virtual assistant provided via a device is to utter a hotword or wake word. With user permission, devices such as smartphones, smart speakers/displays, etc. analyze audio to detect the utterance of a hotword that triggers activation of the virtual assistant. It is important that hotword detection has high precision since not all user speech is directed toward the virtual assistant. Hotword detection, which can involve automatic speech recognition (ASR) and a transcription stack, requires the device microphone to be always on which increases the power demand. The always-on, power-hungry nature of hotword detection is especially challenging on devices with a small form factor or with small batteries, such as hearables (also known as earbuds or smart headphones), smart watches, etc.

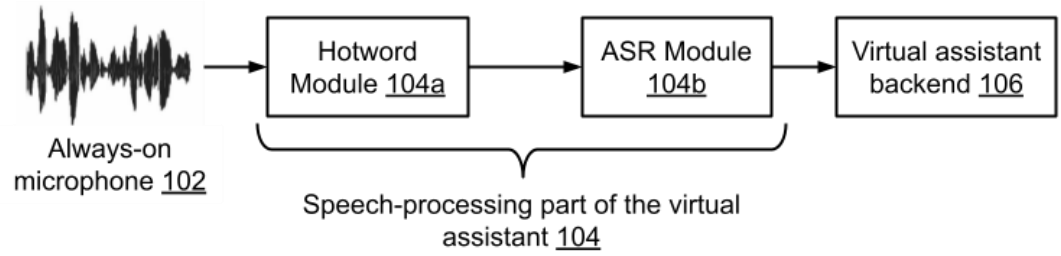

Fig. 1: Conventional hotword detection

Fig. 1 illustrates conventional hotword detection. An always-on microphone (102) detects ambient sound signals (which can potentially include a hotword followed by a command to a virtual assistant) to a speech-processing part (104) of the virtual assistant application. The speech-processing part includes a hotword module (104a) and an ASR module (104b) which are also always on. When a hotword is detected, the hotword and the following command are transmitted to the virtual assistant backend (106) for further action. The power consumption

caused due to the always-on nature of the microphone and of the speech-processing part of the virtual assistant application can be unsustainable on small devices.

DESCRIPTION

This disclosure describes techniques that gate hotword detection with a lightweight neural network that estimates mouth-motion signatures using signals from a low-power, always-on inertial measurement unit (IMU). Because the IMU has lower bandwidth than microphones, signals generated by the IMU can be processed by a neural network that is relatively small, e.g., small enough to reside on the front-end processor of even a low-powered device such as a hearable. Hotword gating can be done without waking up the relatively power hungry digital signal processor (DSP) or system-on-chip (SoC) of the device. IMU-based gating can allow high-precision hotword detection to be achieved at very low power consumption.

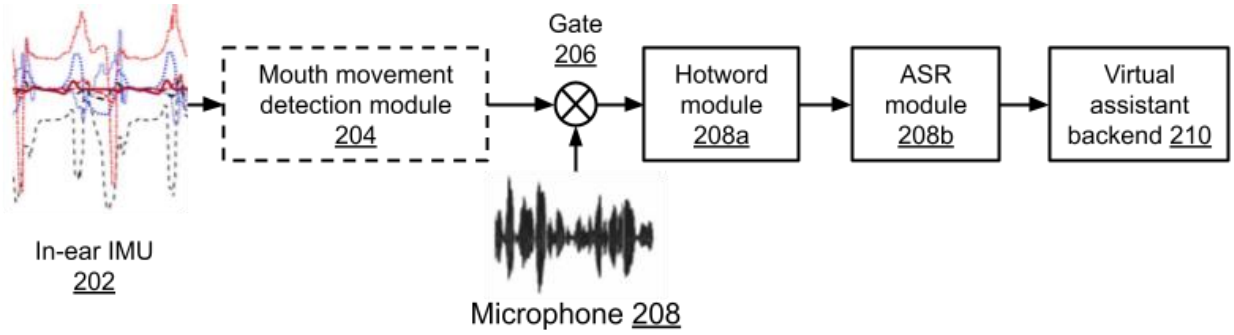

Fig. 2: Mouth movement detection for energy-efficient hotword detection

Fig. 2 illustrates mouth movement detection for energy-efficient hotword detection, per techniques of this disclosure. As in conventional hotword detection, a hotword module (208a) and an ASR module (208b) are fed ambient sounds detected by a microphone (208). Unlike conventional hotword detection, the microphone and the hotword and ASR modules are generally quiescent and consume no power. The modules are woken up only when a mouth

movement detection module (204) unlocks a gate (206) to allow sound waves. The mouth movement detection, explained in greater detail below, is utilized to detect the possible presence of a hotword by analyzing waveforms generated by an on-device IMU (202).

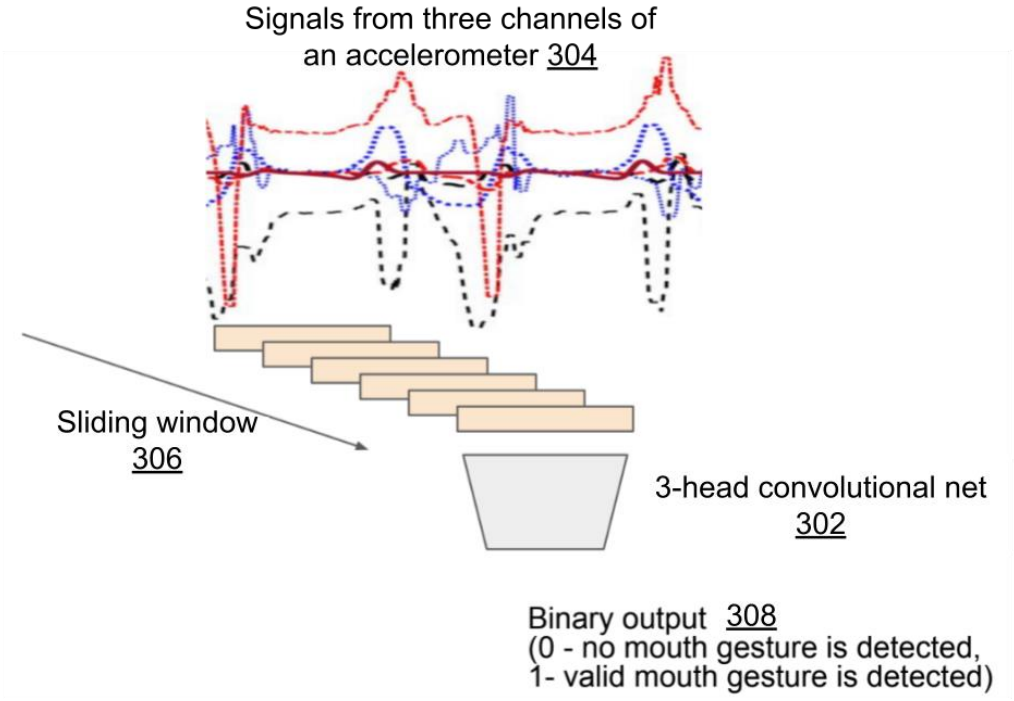

Fig. 3: Mouth movement detection

Fig. 3 illustrates mouth-movement detection. The mouth movement detection module can be a lightweight, multi-headed temporal convolutional network (302), which takes its input across the three channels (304) of a low power accelerometer. The convolutional network can run in a sliding window fashion (306), with the window size being selected to be sufficient to detect a hotword. A window size of the order of 10 ms can detect the signature of a mouth opening gesture that is indicative of the starting point of a hotword. The output (308) of the convolutional neural network indicates if a valid mouth gesture is detected or not, and can be used to gate the hotword, microphone, and ASR modules of the virtual assistant, as described in greater detail below.

Gating of the hotword, microphone, and ASR modules can be done in any of the following ways.

- *Gating just the hotword module but always-on microphone*: If the mouth movement detection network uses relatively long window sizes, e.g., substantially longer than 10 ms, early syllables of the hotword may be missed by the time the microphone is turned on. Therefore, for long window sizes, the microphone can be kept always on while the hotword module is gated, such that computational budget is saved while ensuring precise hotword detection.

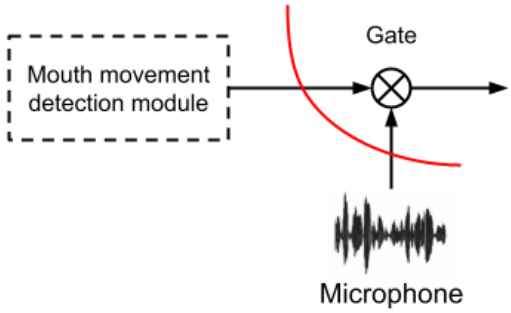

Fig. 4: Using the mouth movement detection module to gate both the microphone and the hotword modules

- *Gating the microphone as well as the hotword module*: If the mouth-movement network uses relatively small window sizes, then, as illustrated in Fig. 4, the hotword module and microphone streaming can be gated. This results in power savings from gating hotword detection computations as well as from gating the microphone. Additionally, turning on the microphone on an as-needed basis can also provide additional benefits, e.g., no hot microphone in the user's space.

The lightweight, low-latency nature of the front-end, mouth movement detection module enables it to be implemented directly on the IMU processing unit. This provides an opportunity to cascade the turning on of the hotword, the ASR, and downstream modules, as the hotword and

upstream modules do not run without mouth movement confirmation from the IMU processor. In this manner, mouth movement based hotword detection, as described herein, can split and turn on the computations performed for hotword detection in an energy-saving, cascaded fashion.

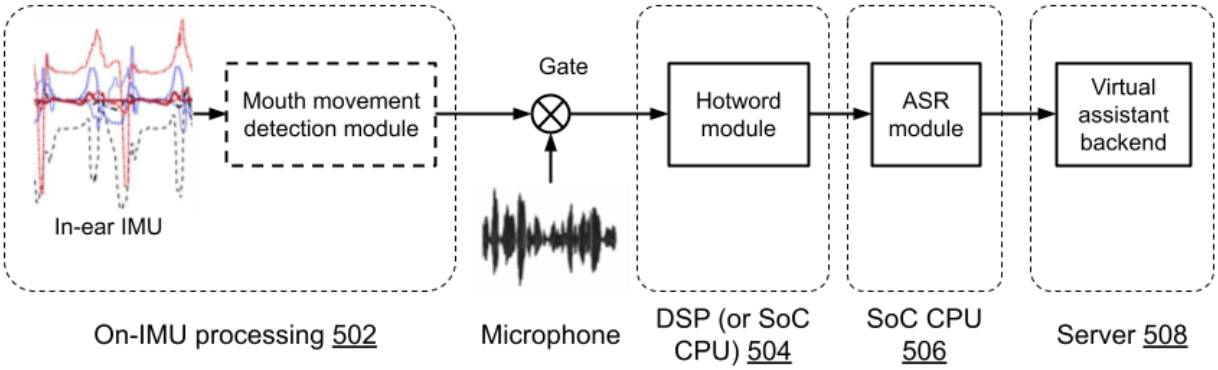

Fig. 5: Cascaded operation of the hotword module, ASR module, and virtual assistant backend, triggered by mouth movement detection

Fig. 5 illustrates cascaded operation of the hotword module, ASR module, and virtual assistant backend, triggered by mouth movement detection. When mouth movement is detected by the mouth movement detection module implemented on the IMU of the device (502), the hotword module, which is located on the digital signal processor (DSP, 504) is turned on. If the hotword is confirmed, the ASR module, located on the SoC CPU (506), is turned on. Subsequently, the virtual assistant backend, located on a server (508), is activated to respond to the query transcribed by the ASR. Cascaded operation of the various modules saves power.

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 the collection of user information (e.g., information about a user's audio, a user's queries, virtual assistant configuration, 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 that gate hotword detection with a lightweight neural network that estimates mouth motion signatures using signals from a low-power, always-on inertial measurement unit (IMU). Because the IMU has lower bandwidth than microphones, signals generated by the IMU can be processed by a neural network that is small enough to reside on the front-end processor of a low-powered device such as a hearable. Hotword gating can be done without waking up power hungry processors of the device. IMU-based gating can allow high-precision hotword detection to be achieved at very low power consumption.

REFERENCES

1. Lee, Jae. "Personalized talking detector for electronic device." U.S. Patent Application 17/551,297, filed December 15, 2021.
2. "Speaker detection by watching lip movements," available online at https://github.com/sachinsdate/lip-movement-net accessed Jun. 28, 2023.
3. Zhang, Li, Parth H. Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. "Accelword: Energy efficient hotword detection through accelerometer." In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 301-315. 2015.