August 2022

Gesture Detection Using Doppler Sonar

D Shin

Gesture Detection Using Doppler Sonar

ABSTRACT

While some computing devices include specialized hardware for gesture detection that enables hands-free operation, many commodity devices such as laptops do not include such hardware. This disclosure describes techniques, implemented with user permission, to automatically detect gestures performed in proximity of commodity devices at no additional cost by employing the on-device speakers and microphone as a Doppler sonar (sound navigation and ranging). The on-device speaker(s) generate an ultrasonic ping signal inaudible to the human ear and the device microphone(s) capture reflections from the user’s body parts. The type of Doppler shift in the reflected signal can indicate the direction of motion of the body part. The signal is provided as input to a trained classifier which can map the detected signal to the type of gesture the user is making. The described techniques can identify gestures quickly, thus enabling a smooth user experience for gesture-based interaction.

KEYWORDS

- Gesture recognition
- Gesture detection
- Doppler effect
- Doppler spectrogram
- Red shift
- Blue shift
- Sonar (Sound Navigation and Ranging)
- Ultrasonic ping
- Ultrasonic pulsing
- Air click
- Air tap
- Radar sensor

BACKGROUND

Some computing devices include functionality to detect various user gestures performed in the vicinity of the device. Such device capabilities enable users to employ gestures to provide commands for the device to perform corresponding actions. For instance, a user can scroll up and down through a document by performing corresponding hand motions in the air in front of the device, invoke a virtual assistant with a hand wave over a device, etc.

User permission is obtained to enable sensors to perform gesture detection. When permitted by the user, detection of user gestures is typically performed in one or more of the following ways: applying computer vision techniques to analyze the feed captured by the device camera, employing dedicated hardware designed for gesture detection (e.g., a radar sensor), using ultrasonic pulsing, etc.

An important requirement is that recognition of the gesture and performance of the corresponding action take place without unnatural lag. The manner in which gesture recognition is implemented on a device impacts latency and thus has an impact on the user experience (UX). For instance, ultrasonic pulsing can be slower than radar-based approaches. Further, approaches that utilize specialized hardware are typically superior for detecting a variety of complex gestures. However, gesture recognition approaches that require devices to be equipped with specialized hardware increase device manufacturing costs and are unavailable on devices that lack the specific hardware components. Moreover, access to gestures detected via such hardware is often protected by the device operating system and may therefore be unavailable for use by many higher-level applications, such as web browsers or other applications.

DESCRIPTION

This disclosure describes techniques to enable user-permitted recognition of various gestures on commodity user devices, such as laptops, smartphones, etc., at no additional hardware cost by employing the on-device speaker(s) and microphone(s) as a Doppler sonar (sound navigation and ranging). Specifically, an on-device speaker generates an ultrasonic ping signal that is inaudible to the human ear and the device microphone captures reflections of the signal from nearby objects, such as the user's body parts in front of the device.

Importantly, reflection of sound waves by a nearby moving target, such as a user's hands engaged in making a gesture, exhibits the Doppler effect, with the pitch increasing as the target moves closer (blueshift) and decreasing when the object moves away (redshift). Therefore, the reflected signal captured by the microphone can indicate the direction of motion of the body part reflecting the signal, which can in turn signify the type of gesture the user is making.

If $s(t) = \cos(2f_{tone}t)$ is the ping signal at the ultrasonic frequency f_{tone}, then the reflected signal can be expressed as $r(t) = A \cos (2 [f_{tone} + v / c] t)$ where A is the signal attenuation scalar, v is the velocity of target (the user's body part), and c is the speed of sound. The velocity is positive when the body part is moving toward the speaker (blue shifting) and negative when the body part is moving away from the speaker (red shifting). For instance, if a user performs an "air tap" by moving their hand toward the device at first and then pulling back, the return signal exhibits a blueshift while the hand is moving toward the device and a redshift when the hand moves backward, away from the device.

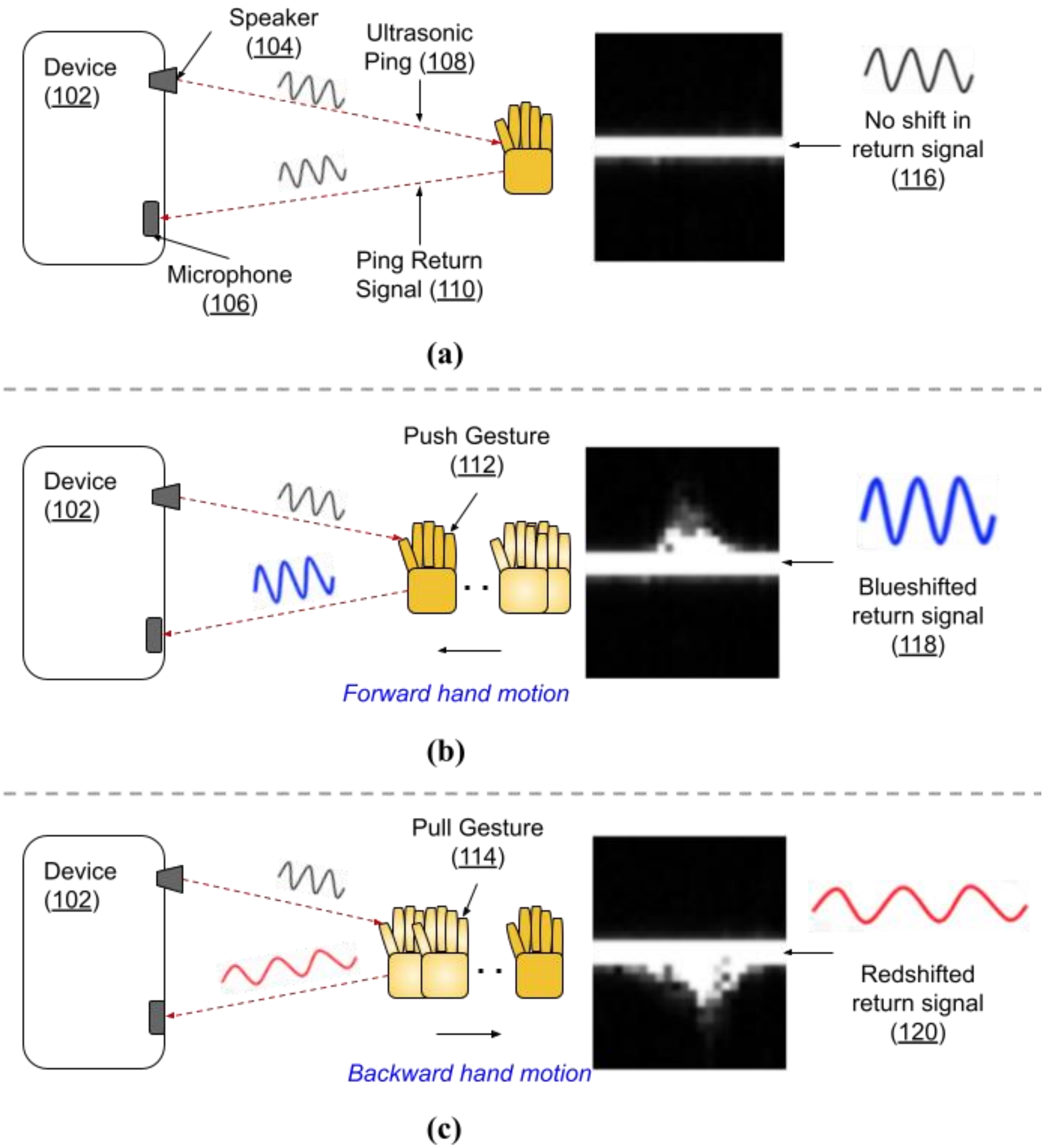

Fig. 1: Using an ultrasonic ping signal to detect gestures performed in front of a device

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. An ultrasonic ping signal (108) is emitted via a speaker (104) of a device (102). The signal is reflected by the user's hand in front of the device and the return signal (110) is captured by the device microphone (106). When the user's hand is stationary (Fig. 1a), there is no shift in the return signal (116). In contrast, a return signal that is blueshifted (118) in comparison to the original ping indicates that the user's hand is pushing (112) toward the device (Fig. 1b) while a redshifted return signal (120) indicates that the hand is "pulling" (114) away from the device.

The reflected signal received by the microphone can be processed using a digital Doppler processing block to yield Doppler spectrogram signatures that can be input to a simple binary classifier that indicates the user's gesture. The Doppler processing includes:

1. **Mixing** (optional): Ultrasonic signal mixing can isolate the Doppler frequency s(t) * r(t), i.e., the difference frequency between the ping source and reflection. Operationally, it generates two frequency components: the desired Doppler component and a high-frequency ghost component. The high-frequency component can be removed easily using a simple low pass filter (LPF).
2. **Windowing**: This windowing operation is performed to suppress side lobes in frequency to increase the Doppler spectral peak resolution that intersects with the peak of the direct current (DC) signal. The operation can use a typical window such as Hann.
3. **Transforming**: A Fast Fourier Transform (FFT) can be applied to convert the Doppler signal into frequency domain. Correct application of all steps up to this point yields an almost pure tone at the frequency corresponding to the velocity of the moving target in free air.
4. **Stacking**: To enable robust classification, stacking is employed to buffer the instantaneous gestures for a duration of sufficient length, such as tens of milliseconds. The buffers can be deemed as Doppler spectrograms. Distinct spectrogram signatures can be associated with distinct gestures, depending on the blue and red shift patterns within the spectrogram. For instance, as shown in Fig. 1, "push" and "pull" gestures with the hand result in opposite

spectral patterns. An “air tap” is basically a pattern of a “push” followed by that of a “pull.” In contrast, a “tickle” gesture made by tickling fingers quickly in free air leads to bursts of patterns of opposing movement that indicate movement that rapidly switches directions back and forth.

A collection of such spectrograms can be used to train a classifier, e.g., implemented using a neural network with an application-specific architecture, or other suitable techniques. Fig. 2 shows some examples of gestures and corresponding Doppler spectrograms. An “air click” is detected when the user quickly taps the vacuum above the device speaker using the palm while a “tickle” is detected when a tickling gesture is made in free air. For “air click” the clear negative dip followed by a positive dip around the DC band corresponds to the velocity of the user’s palm coming close to the device then moving away. For “tickle” the near-random mini-bursts of negative and positive dopplers represent fingers moving rapidly.

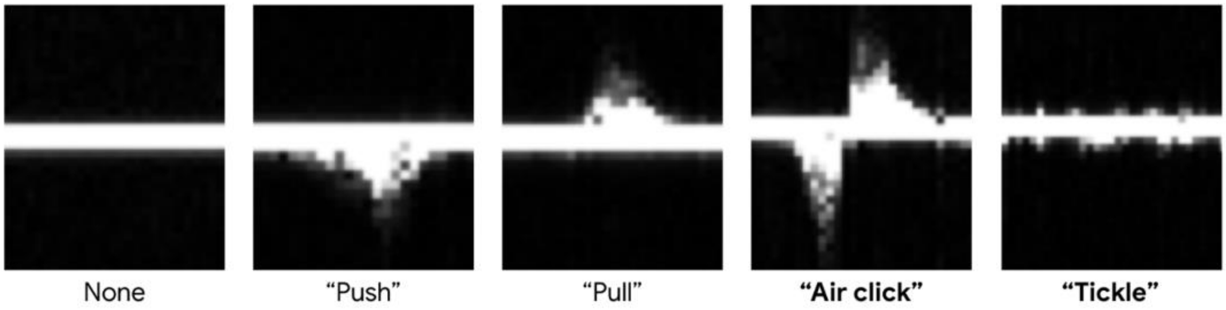

Fig. 2: Example spectrograms and corresponding gestures

Doppler spectrograms generated in an operational implementation as described above are input to the trained classifier. The output of the classifier can indicate the gesture the user is performing. Unlike the latency that users may experience when gestures are recognized using ultrasonic pulsing, the techniques described herein can identify gestures in under a second, thus providing detection speeds quick enough for a smooth user experience (UX) for gesture-based interactions within close proximity of the device.

To avoid excessive power consumption, ultrasonic pings can be generated at suitable intervals that are set by the developers and/or determined dynamically at runtime. Similarly, the frequency and volume of the ultrasonic ping signal can be specified by the developers and/or set dynamically at runtime. For instance, higher frequencies can be used if a greater sampling rate is necessary.

In contrast to radar-based approaches that require specialized hardware, the techniques described in this disclosure can be used to enable gesture recognition on any device that includes a built-in speaker and microphone, making it possible to provide gesture-based interaction on legacy devices and without incurring additional hardware costs. However, compared to radar-based approaches, only a small set of gestures at relatively low resolution of motion can be detected via the described techniques. If available on the device, two (or more) microphones with suitable baseline separation can be employed. Availability of such hardware can expand the set of gestures that can be recognized with the described techniques. In such a case, standard delay filter or FFT-phase methods can be employed to form a beam of the ping signal to estimate the angle-of-arrival (AoA) before the mixing step described above. The spectrogram can be composed of the {return signal, angle} tuple over time, thus providing the ability to detect motion along two axes.

The operative range of the gesture detection that can be performed using the described techniques can be lengthened by using a Doppler configuration in transmission mode instead of reflection mode. In such a setup, the user can perform gestures while holding or wearing a mobile device, such as a smartphone, smartwatch, etc., that is used for tone generation. The gesture can then be recognized in under a second on a separate device, such as a laptop, that captures the tone and performs gesture recognition by processing it in the same manner as that described above for reflection mode. For instance, a user wearing a smartwatch capable of generating the ultrasonic ping tone can shake their wrist as a command to log out of the laptop. Similarly, a user can air tap in front of the laptop while holding a smartphone that generates the appropriate ultrasonic signal to pause the currently playing audio.

The described gesture detection techniques can be implemented as part of a device operating system, or with appropriate user permission, as part of any application, e.g., a web browser or other application. The detected gestures can be mapped to commands for performing relevant actions such as scrolling text, zooming images, closing windows, playing and pausing media content, muting sound, turning camera off, etc. For instance, implementation of the techniques can enable a user to activate a virtual assistant with a simple wave of the hand over the device. In transmission mode, the techniques can additionally be employed for measuring proximity between two devices based on the Doppler cues between the device used to generate the tone and the device that receives it.

Further to the descriptions above, a user is 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 gestures, commands, a user’s devices, or a user’s preferences), and if the user is sent content or communications from a server. In addition, certain data are 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. Thus, the user has control over whether and 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, implemented with user permission, to automatically detect gestures performed in proximity of commodity devices at no additional cost by employing the on-device speakers and microphone as a Doppler sonar (sound navigation and ranging). The on-device speaker(s) generate an ultrasonic ping signal inaudible to the human ear and the device microphone(s) capture reflections from the user's body parts. The type of Doppler shift in the reflected signal can indicate the direction of motion of the body part. The signal is provided as input to a trained classifier which can map the detected signal to the type of gesture the user is making. The described techniques can identify gestures quickly, thus enabling a smooth user experience for gesture-based interaction.

REFERENCES

1. Gupta, Sidhant, Daniel Morris, Shwetak Patel, and Desney Tan. "Soundwave: using the Doppler effect to sense gestures." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1911-1914. 2012.
2. Weiss, Lior. "Hand Gesture Recognition with Wi-Fi Doppler Imaging." Available online at https://www.celeno.com/blog/hand-gesture-recognition-wi-fi-doppler-imaging accessed 22 May 2022.
3. Shin, D, "Sleep Mode Activation Based on Detecting Laptop Lid State via Ultrasonic Ping Signal", Technical Disclosure Commons, (June 16, 2022) https://www.tdcommons.org/dpubs_series/5211