May 2021

Camera-driven Beamforming for Better Audio Capture

Chris Igo

Camera-driven Beamforming for Better Audio Capture

ABSTRACT

Background noise picked up by microphones can be mitigated using audio beamforming. However, if a user loses alignment with the beam, the user’s speech can be severely attenuated or lost in the noise. This disclosure utilizes a camera, with user permission, to obtain coordinate information for a user that is providing audio input, e.g., in a voice/video call, during an audio recording, etc. The information is utilized to dynamically adjust the direction of the beam of an audio beamformer such that a user’s speech, e.g., as received by the beamformer is consistent regardless of the current position or movement of the user.

KEYWORDS

- Audio beamforming
- Speech capture
- Sensor fusion
- Noise cancelation
- Directional cancelation
- Noise enhancement
- Beam misdirection
- Audio fidelity

BACKGROUND

Background noise picked up by microphones can be mitigated using audio beamforming, which is a technique that creates a focused, directional beam that isolates the user’s speech from surrounding disturbances. However, if a user loses alignment with the beam, e.g., if the user moves or if the beam otherwise points away from the user, the user’s speech as detected can be severely attenuated or even lost in the noise. Although an audio beamformer generally delivers a high signal-to-noise ratio, its performance is tightly coupled to the direction of its beam, e.g., it can lack robustness.

DESCRIPTION

This disclosure uses coordinate information from a camera, generated by application of machine learning techniques, to dynamically adjust the direction of the beam of an audio beamformer such that the user's speech, e.g., as received by the beamformer and as transmitted to a far end of a phone call (or utilized in other situations such as a video call, recording, etc.) is consistent regardless of the position or movement of the user around the microphone. The user is provided with indications that the camera may be used for such a purpose and options to enable or disable such use.

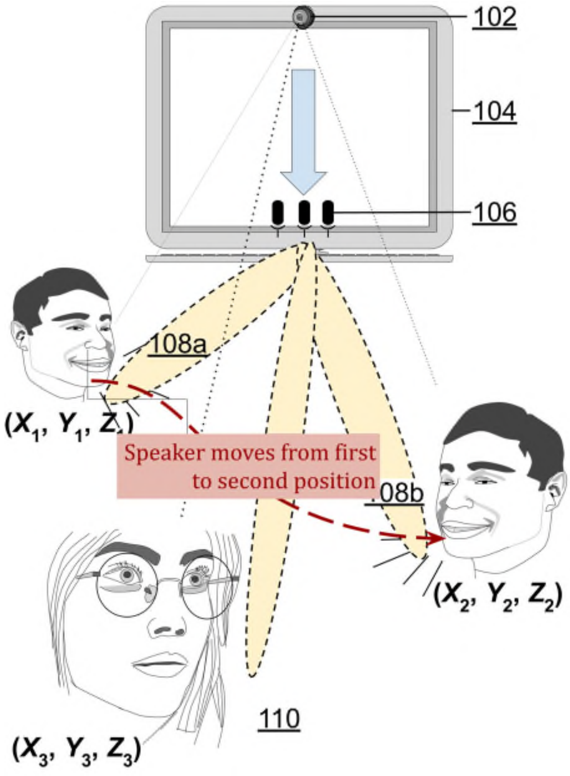

Fig. 1: Camera-driven audio beamforming

Fig. 1 illustrates camera-driven audio beamforming, per the techniques of this disclosure. A camera (102), and a microphone array (106) that includes one or more microphones, are embedded in a device (104), e.g., laptop, desktop, smartphone, smart-home device, etc. A human speaker is at a position (X_1, Y_1, Z_1). With specific user permission, the camera is utilized to detect the coordinates of the speaker, which can include distance, angle, and size, using, e.g., a face detector.

For typical human movements, e.g., in a home, office, or conference room setting, the coordinates of the speaker can be updated twice a second, e.g., at 0.5 Hertz. The detected coordinates are communicated to the microphone array, which uses the location information of the speaker to direct an audio beam (108a) towards the speaker. After some time, the speaker moves to a new position (X_2, Y_2, Z_2). The camera is utilized to determine the new coordinates of the speaker which are communicated to the microphone array. The information is utilized to redirect the audio beam (108b) towards the new location of the speaker.

If a different speaker located at (X_3, Y_3, Z_3) starts speaking, the camera determines the coordinates of the different speaker and communicates their coordinates to the microphone array, which uses the location information of the different speaker to redirect the audio beam (110) towards the different speaker.

Some example use cases are described below.

Example: Camera-triggered audio beamforming

In this use case, camera-driven audio beamforming detects a temporary absence of the user and disables beamforming such that noise enhancement due to the beam misdirection is forestalled. In an initial state, the user is on a video call, and audio beamforming actively points a beam towards the user. In a next state, the user steps away, e.g., to get a cup of coffee. Camera-driven audio beamforming recognizes that the user is out of the beam field and disables beamforming, enabling the user to seamlessly continue their conversation without losing audio fidelity. In a subsequent state, the user returns to the front of their computer to continue the call, which causes re-enablement of beamforming.

Example: Accurate beamforming

Although adaptive beamforming can be done solely with microphones, the incorporation of location data provided by the camera can enable accurate and energy-efficient, pointing of the beam to the person currently speaking, adapting dynamically as the person moves.

Example: Turning off and on beamforming in the presence of multiple speakers

As explained previously, beamforming works optimally in a single direction, e.g., to a single target, such that when the number or directions of targets increase, beamforming can degrade due to misdirection on a silent user, due to attenuation of a speaking user, due to noise amplification, etc. Per camera-directed beamforming, the presence of multiple people in the camera frame can cause the microphone array to appropriately defocus (e.g., make more diffuse or less sharp), the beam such that the multiple persons in the frame are accommodated. When many persons (relative to the number of microphone sensors) are in the frame or are determined to be recent speakers, beamforming can be turned off, e.g., the beam can become uniform (omnidirectional), such that all participants can participate in the call without difficulty.

The described techniques can be used in a device, e.g., laptop, desktop, smartphone, tablet, smart-home device, smart speaker, smart display, virtual assistant, audio/video conferencing hardware, etc., that includes a camera and one or more microphones, Example applications include audio or video calls, conference calls, public-address equipment, audio recording, etc.

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 the collection of user information (e.g., information about a user’s preferences), and if the user is sent content or communications from a server. In addition, certain data is 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 is treated so that no personally identifiable information can be determined for the user. Thus, the user has 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 utilizes a camera, with user permission, to obtain coordinate information for a user that is providing audio input, e.g., in a voice/video call, during an audio recording, etc. The information is utilized to dynamically adjust the direction of the beam of an audio beamformer such that a user’s speech, e.g., as received by the beamformer is consistent regardless of the current position or movement of the user.

REFERENCES

[1] Rudberg, Tore, and Christian Schuldt. “Controlling focus of audio signals on speaker during videoconference.” U.S. Patent 10,356,362, issued July 16, 2019.