January 2023

Spatially Realistic Audio in a Video Conference Based on User Head Orientation

D Shin

Jian Guo

Spatially Realistic Audio in a Video Conference Based on User Head Orientation

ABSTRACT

In current video conferencing applications, the audio of the speech of a participant is captured without any indication of the spatial positioning or body orientation of the speaker in relation to a device camera used to capture the corresponding video. Therefore, the audio experience in video conferencing lacks spatial and directional richness. This disclosure describes techniques to enhance the spatial richness of the audio in a video conference based on a user's head orientation. With user permission, head orientation is estimated using measurements from device sensors of earbuds or another device used by a video conference participant. Head orientation measurements for the participants are used to apply appropriate positional correction to the audio using a head-related transfer function (HRTF). Implementation of the techniques can improve the spatial accuracy of the audio feed within a video conference, thus making the conversations sound more realistic.

KEYWORDS

- Video conferencing
- Head orientation
- Head pose
- Head-Related Transfer Function (HRTF)
- Spatial audio
- Immersive audio
- Audio correction
- Earbuds

BACKGROUND

Devices such as laptops, smartphones, etc. are commonly used to engage in video conferencing. During a video conference, device microphones capture the user's speech which is relayed to the other participants. In current video conferencing applications, the raw audio of a user's speech of is relayed with a direct mono-to-stereo mapping such that the other participants receive the same raw audio repeated through the left and right channel of the audio output delivered via the device speakers or other audio device, such as headphones, earbuds, etc. The audio of the speech is captured without any indication of the spatial positioning or body orientation of the speaker in relation to the device camera that is used to capture the corresponding video during the video conference. As a result, the audio experience in video conferencing lacks spatial and directional richness, thus coming across as flat.

DESCRIPTION

This disclosure describes techniques to enhance the spatial richness of the audio signal in a video conference based on estimating the user's head orientation. With appropriate permissions from a videoconference participant, the head orientation of the participant is estimated with three degrees of freedom (3DOF) by measuring three-dimensional acceleration with the gyroscope in Inertial Measurement Unit (IMU) sensors contained within the participant's earbuds or other head-worn device. The 3DOF head orientation measurements for the speaking and the listening participants within a video conference can be employed to determine the speaking and listening angles of the respective participant in relation to the device camera, which is typically present near the top of the screen of the device via which the participant joins the video conference.

The combined 6DOF head orientation information (i.e., 3DOF for the speaker and 3DOF for the listener) can then be used to apply appropriate positional correction to the captured audio

of the speaker being output on the side of the listener. With permission, the correction can be achieved by modifying a parameterized Head-Related Transfer Function (HRTF) which can, for instance, be a lookup table with six parameters.

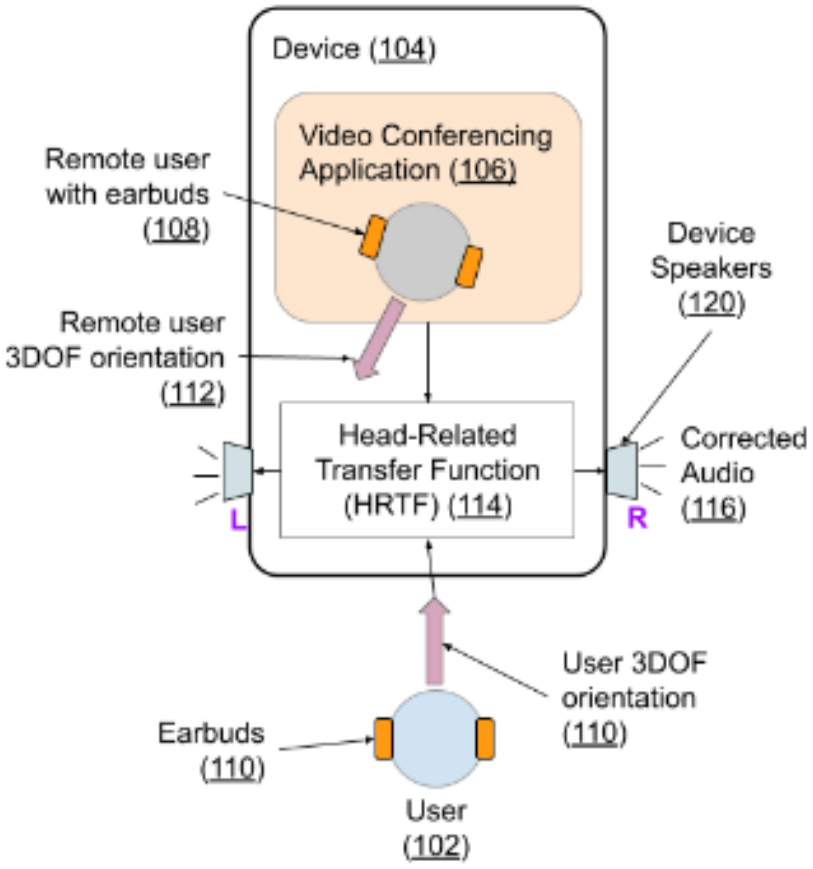

Fig. 1: Applying appropriate orientation correction to the audio feed in a video conference

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. As shown in Fig. 1, a user (102) is in a video conference with a remote user (108) via a video conferencing application (106) on the user's device (104). With appropriate permission, the 3DOF head orientations of the two users in the video conference (110 and 112, respectively) are obtained locally at each end via the IMUs within their respective earbuds (110) through a traditional approach, e.g., a Mahony-type filter. The two respective 3DOF

measurements can be used to look up the appropriate HRTF (114). The HRTF can be calibrated offline using anechoic chamber measurements.

The audio can be spatially corrected (116) according to the HRTF. For instance, raw audio *a* of the speaker can be spatially corrected for the 3DOF orientations of the speaker *[3DOF(s)]* and the listener *[3DOF(l)]* by applying the following convolutional equations for the Left and the Right audio channel, respectively:

```
a_Left = HRTF_Left({3DOF(s), 3DOF(l)}) * a
a_Right = HRTF_Right({3DOF(s), 3DOF(l)}) * a
```

In the above equations, HRTF_Left and HRTF_Right denote the respective HRTF scan results for the left and right audio channels for various orientations of the source of the audio. When the head orientation of a user shifts, the HRTF is updated and corresponding adjustments are made to the audio based on the updated HRTF for the new orientation.

The spatially corrected audio for the left and the right channels can be played back via the respective speakers (120) of the device (marked in Fig. 1 as L and R, respectively). Alternatively, or in addition, the corrected audio can be sent directly to the respective channel for the left and right earbud of the participants.

The techniques described in this disclosure can be implemented within any video conferencing application and can support any earbuds or other devices that contain sensors that enable determination of head orientation. Although the above description illustrates a video conference between two users, the techniques can support any number of participants as long as head orientation measurements can be obtained. Implementation of the techniques can improve the spatial accuracy of the audio feed within a video conference, making conversations sound more realistic.

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 video conferences, head orientation, audio input and output devices, 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 to enhance the spatial richness of the audio in a video conference based on a user's head orientation. With user permission, head orientation is estimated using measurements from device sensors of earbuds or another device used by a video conference participant. Head orientation measurements for the participants are used to apply appropriate positional correction to the audio using a head-related transfer function (HRTF). Implementation of the techniques can improve the spatial accuracy of the audio feed within a video conference, thus making the conversations sound more realistic.