November 2022

Participant-Targeted Spatial Audio Generation Using Gaze Matching Over a Multi-Grid Video Conferencing Setup

D. Shin

PARTICIPANT-TARGETED SPATIAL AUDIO GENERATION USING GAZE MATCHING OVER A MULTI-GRID VIDEO CONFERENCING SETUP

Introduction

In a traditional video conferencing setup with multiple participants, it is hard for a speaker to naturally indicate to whom the speaker is directing their speech. This is because, unlike in real-world interactions, the acoustics of the speaker's voice remain constant regardless of whom the speaker is directing their gaze towards. As a result, despite improvements in video compression, high-quality audio, and low-latency streaming, video conferencing is still not as immersive as real-world conversations.

Summary

The present disclosure relates to systems and methods to provide a more immersive video conferencing experience. More particularly, aspects of the present disclosure provide a method for determining one or more participants on a video conference call to which a speaking participant is directing their speech towards by tracking the speaker's gaze. In response to determining which participants the speaker is directing their speech towards, one or more parameters of the video conference may be modified for one or more participants. For example, the one or more parameters could include parameters for spatial audio modulation (e.g., amplitude and/or reverberation effect) that may be applied to the speaking participant's voice. Thus, the present disclosure can provide a more immersive video conferencing experience. Additionally, some implementations of the present disclosure can be used to more efficiently allocate network resources (e.g., bandwidth). For example, the one or more parameters could include a quality setting of the video and/or audio streams.

Detailed Description

The present disclosure is generally directed towards systems and methods to increase immersion for video conference participants. More particularly, aspects of the present disclosure include a computer-implemented method for immersive video conferencing. The method can include displaying a plurality of display elements corresponding to a plurality of participants on a display device. The method can also include obtaining information indicative of a speaking participant's gaze. The method can further include determining a target display element from the plurality of display elements (i.e., a display element which the speaking participant is focused on) based on the participant's gaze. Additionally, the method can include providing information indicating which of the plurality of listening participants the speaking participant is focused on based on the target display element.

In some implementations, obtaining information indicative of a speaking participant's gaze can be accomplished by a gaze matching engine. For example, the gaze matching engine may include machine-learned model(s) trained to determine a participant's gaze (e.g., a gaze vector, a gaze direction, a gaze target, etc.). The gaze matching engine can receive one or more image frames of a speaking participant from a video capture device (e.g., a webcam). The gaze matching engine can determine a pair of face coordinates and gaze coordinates for one or more of the image frames. The gaze matching engine can compare the pair of face coordinates and gaze coordinates to a plurality of calibration pairs. The gaze matching engine can determine which of the plurality of calibration pairs is closest to the pair of face coordinates and gaze coordinates.

Other aspects of the present disclosure provide a method for receiving information indicating which of a plurality of listening participants a speaking participant's gaze is directed.

In response to the received information, a computing system can modify various aspects of a stream. For example, the computing system may modify a volume of the speaking participant, a reverberation effect of the speaking participant, a quality of an audio and/or video stream, etc.

Figure 1 depicts an example computing system 100 in which systems and methods in accordance with the present disclosure can be executed. The computing system 100 includes a speaker computing device 102. The speaker computing device 102 contains one or more processors 104 and memory 106, which may contain data 108 and instructions 110 configured to carry out the methods disclosed herein. The speaker computing device 102 can further include a gaze capture device 112, an audio capture device 114 (e.g., microphone), a video capture device 116 (e.g., webcam), and a display device 118 (e.g., computer monitor, touch screen, television, etc.). The gaze capture device 112 can provide information indicating which region of the display device 118 a participant's gaze is directed. In some implementations, the video capture device 116 may include, or otherwise be communicatively coupled to the gaze capture device 112. In other implementations, the gaze capture device 112 may be a device distinct from the video capture device 116. For example, the gaze capture device 112 may be an augmented reality (AR) or virtual reality (VR) device (e.g., a headset, etc.) with gaze tracking capabilities.

The computing system 100 further includes a server computing system 122 and a listener computing device 132. The server computing system can include processor(s) 124 and memory 126, which can include data 128 and instructions 130 as previously described with regards to speaker computing device 102. The server computing system 122 may relay video and/or audio streams between one or more speaker computing devices 102 and one or more listener computing devices 132. The server computing system 122 may modify parameters of the video and/or audio streams in accordance with aspects of the present disclosure. Additionally, or

alternatively, the server computing system 122 may relay information indicating where a participants gaze is directed in accordance with aspects of the present disclosure.

The listener computing device 132 can include one or more processors 134, memory 136 (containing data 138 and instructions 140) as described with regards to the speaker computing device 102. The listener computing device 132 can further include an audio playback device 142 (e.g., a speaker, etc.) and a video capture device 144 (e.g., a webcam, network of cameras sufficient to generate a three-dimensional representation, etc.).

In some implementations, the speaker computing device 102 and the listener computing device 132 may communicate directly over a network 120 (e.g., LAN, WAN). Additionally, or alternatively, the speaker computing device 102 and the listener computing device 132 may communicate over the network 120 with a server computing system 122. For example, the server computing system 122 may host a video conferencing session that facilitates communication between the speaker computing device 102 and the listener computing device 132.

Figure 2 depicts an example implementation 200 for obtaining calibration data to allow the video capture device 116 to function as the gaze capture device 112 using a gaze matching engine. A speaking participant 202 is positioned within a field of view 214 of the video capture device 116. The display device 118 is configured to display a grid (e.g., a 3 x 4 grid) of cells 206. A dot 208 is configured to be overlaid over each of the cells 206 sequentially. As the dot 208 is overlaid on each of the cells 206, data depicting the speaking participant 202 is captured via the video capture device 116.

For each of the images, a pair of face coordinates 216 and a gaze vector 210 is extracted. The face coordinates 216 can indicate a region within the field of view 214 at which the speaking

participant's face is positioned. The face coordinates 216 may be determined using a face detector (e.g., a machine-learned model configured to detect the position of a face, etc.). The gaze vector 210 corresponds to the direction that the speaking participant's eyes are looking. The gaze vector 210 may be determined using a gaze extractor (e.g., a machine-learned model configured to detect a gaze, etc.). The calibration data may be represented in the form $\{(x, y), (x', y')\}_i$, where (x, y) represents the face coordinates 216 and (x', y') represents the gaze vector 210 determined for each calibration image i.

Figure 3 depicts an example runtime implementation 300 according to aspects of the present disclosure. As depicted, the speaking participant 202 is positioned within the field of view 214 of the video capture device 116. The display device 118 is configured to display the grid of cells 206. Each of the cells 206 can depict a video stream corresponding to a listening participant 310 or a group of listening participants 312, received via network 120 from the video capture device 144 of one or more listener computing devices 132. Additionally, or alternatively, each cell 206 can include a corresponding depiction (e.g., a profile picture, icon, etc.) and/or the names of one or more listening participants.

The gaze detection module 302 determines which cell 206 the speaking participant 202 is directing their gaze (e.g., a target cell 308). In an example implementation, the gaze detection module utilizes the video capture device 116 to determine which cell 206 the speaking participant 202 is directing their gaze. For each frame captured by the video capture device 116 during runtime, a pair of face coordinates 216 and a gaze vector 210 is extracted. The gaze matching engine 304 compares the runtime pair to predetermined calibration data 314 to determine which of the cells 206 the speaking participant's 202 gaze is directed. The predetermined calibration data 314 contains pairs of face coordinates 216 and gaze vectors 210

for one or more calibration images corresponding to cells 206. In one implementation, the gaze matching engine determines which of the calibration images best describes the current frame via L2-minimization. For example, to select a calibration image, the gaze matching engine may determine:

$$\text{argmin(i)} \parallel \{(x, y), (x', y')\}_i - \{(x, y), (x', y')\}_{\text{runtime}} \parallel^2$$

where $\{(x, y), (x', y')\}_i$ corresponds to a pair of face coordinates 216 and gaze vector 210, respectively, for each calibration image i; $\{(x, y), (x', y')\}_{\text{runtime}}$ corresponds to a pair of face coordinates 216 and gaze vector 210, respectively, for the runtime image; and argmin(i) determines which calibration image i results in the lowest value.

In another implementation, the gaze detection module 302 utilizes a gaze capture device 112 to determine which cell 206 the speaking participant 202 is directing their gaze. For example, the gaze capture device 112 may be an AR/VR device with gaze tracking capabilities.

In some implementations, the output of the gaze detection module 302 can be passed through an attention filter 316. The attention filter 316 (e.g., a low-pass attention filter, etc.) can be used to process the output of the gaze detection module 302, which may, in some instances, be jittery due to misdetections or false cell detections. Specifically, attention filter 316 preserves sharp attention changes (e.g., moving from talking to one listening participant to another), while minimizing jitter when focusing on one target cell 308.

The output of the gaze detection module 302—or in some implementations, the attention filter 316—indicates a target cell 308 that the speaking participant's 202 gaze is directed. The listening participant 310 or group of listening participants 312 located in the target cell 308 become the target participant(s). The speaker computing device 102, server computing device

122, and/or listener computing device 132 may be configured to modify one or more parameters associated with the video conference call based on the target participants(s).

In one example implementation, the listener computing device 132 may receive, via network communication 120, from the speaker computing device 102 and/or server computing system 122, a video stream of speaking participant 202 from video capture device 116, an audio stream of speaking participant 202 from the audio capture device 114, and information indicating whether a listening participant 322 is the target participant or in a group of target participants. In response to the information indicating whether the listening participant 322 is a/the target participant, the listener computing device may be configured to modify the audio stream of the speaking participant 202 with a spatial audio modulator 318.

The spatial audio modulator 318 may modify the amplitude and/or reverberation of the audio stream. For example, if the listening participant 322 is a/the target participant, the amplitude of the audio stream of the speaking participant 202 may be increased and/or the reverberation effect decreased. On the other hand, if the listening participant 322 is not a/the target participant, the amplitude of the audio stream of the speaking participant 202 may be decreased and/or the reverberation effect increased. An audio playback device 142 plays the modulated audio stream for the listening participant 322.

In another example implementation, in response to the information indicating whether the listening participant 322 is a/the target participant, the server computing system 122 may modify the bitrate of the audio and/or video streams of the speaking participant 202 being transmitted over the network 120 to the listener computing device 132 used by the listening participant 322. For example, if the listening participant 322 is not a/the target participant, the server computing system 122 may decrease the bitrate of the audio and/or video streams of the speaking participant

202 transmitted to listener computing device 132 to save bandwidth. Additionally, or alternatively, the speaker computing device may decrease the bitrate of the audio and/or video streams of the speaking participant 202 transmitted to the listener computing device 132 over the network 120 if the listening participant 322 is not a/the target participant.

Figures

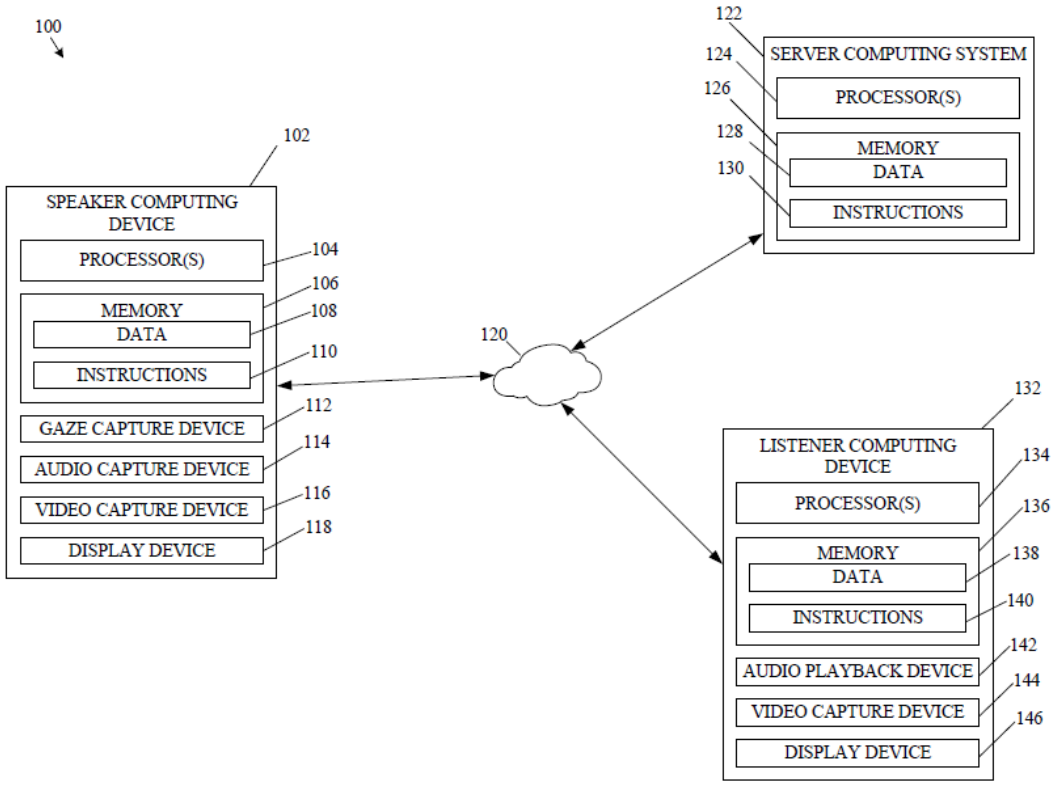

FIG. 1

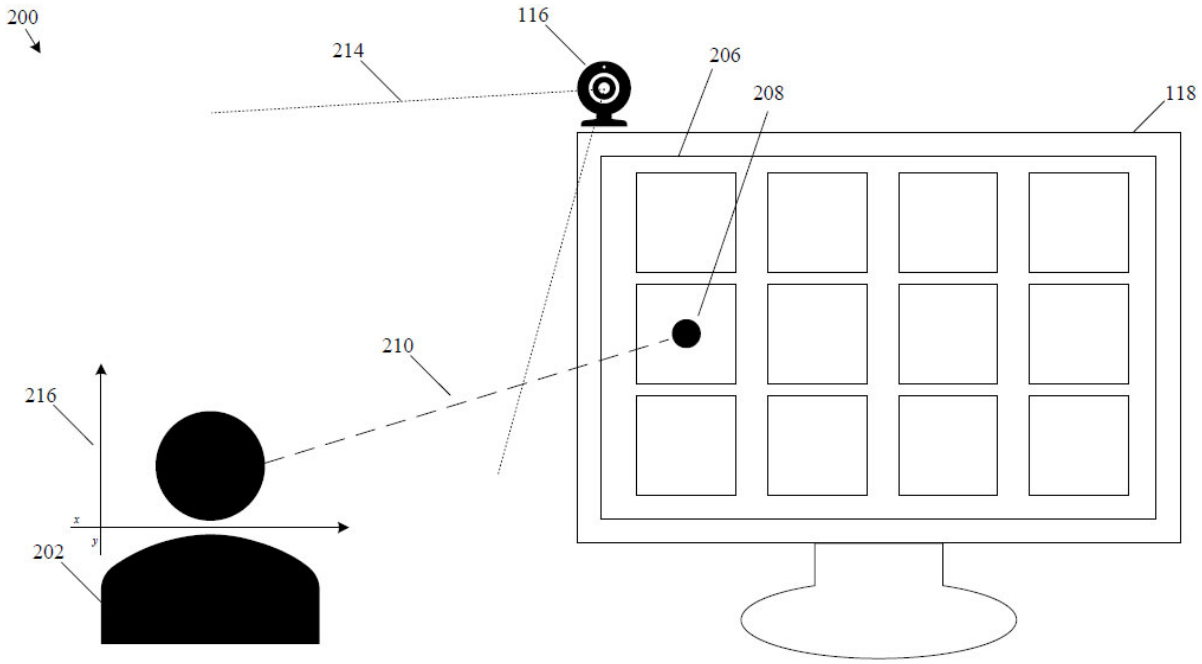
200
116
214
206
208
118
210
216
202
x
y

FIG. 2

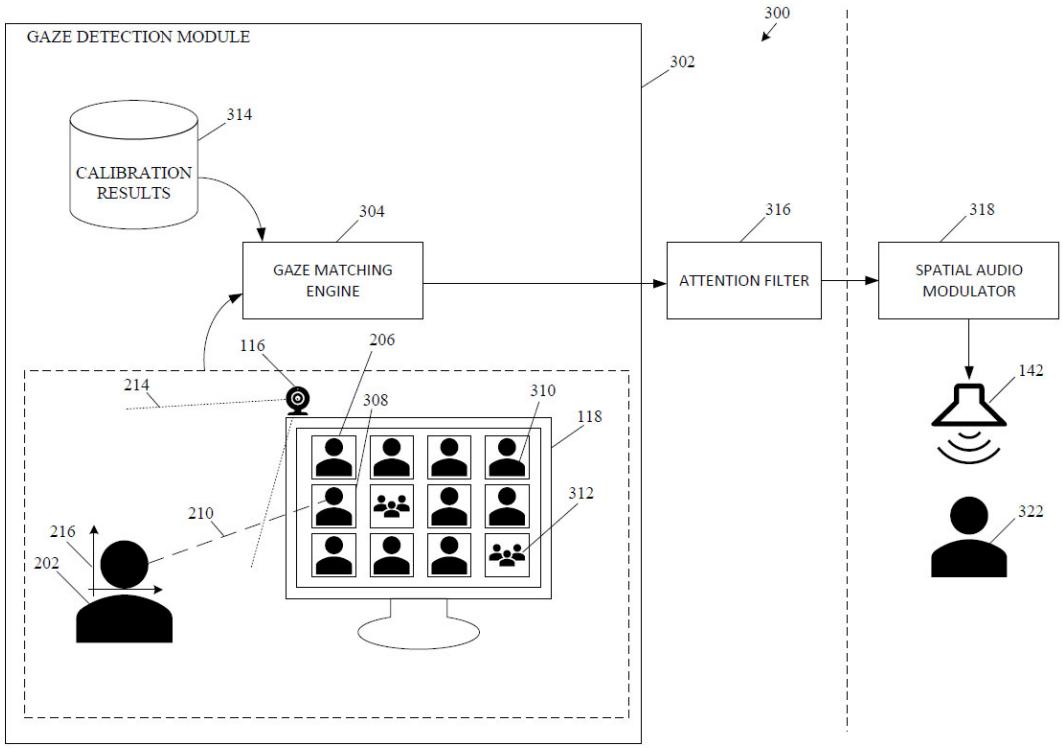
300
GAZE DETECTION MODULE
302
314
CALIBRATION RESULTS
304
GAZE MATCHING ENGINE
316
ATTENTION FILTER
318
SPATIAL AUDIO MODULATOR
116
206
214
308
310
118
312
210
216
202
142
322

FIG. 3

Abstract

The present disclosure is generally directed towards systems and methods to provide a more immersive video conferencing experience. More particularly, aspects of the present disclosure include a computer-implemented method for immersive video conferencing. The method can include displaying a plurality of display elements corresponding to a plurality of listening participants on a display device. The method can also include obtaining information indicative of a speaking participant's gaze. The method can further include determining which of the plurality of display elements the speaking participant is focused on based on the speaking participant's gaze. Additionally, the method can provide information indicating which of the plurality of listening participants the speaking participant is focused on. A computing system can modify, in response to the information, at least one of: a volume of the speaking participant, a reverberation effect of the speaking participant, or a quality of an audio and/or video stream.