August 2023

Automated Audio Realism Detection

Mark Saddler
Meta Platforms Technologies, LLC

Morteza Khaleghimeybodi
Meta Platforms Technologies, LLC

Antje Ihlefeld
Meta Platforms Technologies, LLC

AUTOMATED AUDIO REALISM DETECTION

Inventors:
Antje Ihlefeld
Mark Saddler
Morteza Khaleghimeybodi

FIELD OF THE INVENTION

[0001] The present disclosure generally relates to detecting realism of presented content, and specifically relates to automated audio realism detection.

BACKGROUND

[0002] Artificial reality systems present virtual content to users. The virtual content may include images and/or audio. Note that the audio is generally customized for the viewing user to increase a perceived level of realness of the experience (to ensure the immersive audio contents that are rendered for the user are perceptually indistinguishable from reality). However, there may be mistakes or imperfection in the customized audio that degrade the perceived level of realness. Conventionally, realism of an artificial reality experience is measured via questionnaires to determine the perceived level of realism, and this feedback may be used to adjust the customized audio in an attempt to improve the perceived level of realism. Note that such a process requires active participation of the user (prompting the user to provide feedback) which can degrade the experience.

Detailed Description

[0003] Described herein is automated audio realism detection. An artificial reality system ("system") may perform the automated audio realism detection system. The system may include, e.g., a display assembly, an eye tracking system, a speaker and microphone arrays, and a

controller. In some embodiments, the system may also include one or more electrodes (e.g., to measure eye movement). The system may integrated into, e.g., a headset (e.g., glasses form factor). In some embodiments, portions of the system may additionally be integrated into in-ear devices (IEDs).

[0004] The display assembly presents virtual content to a user of the system. The display includes one or more display elements that present the virtual content. The display elements may be integrated into a headset. The display element may be, e.g., a waveguide display, a liquid crystal display, or some other display able to present visual content to the user. Note that in some embodiments, some or all of the display elements are opaque and do not transmit light from a local area. For example, the local area may be a room that a user wearing the headset is inside, or the user wearing the headset may be outside and the local area is an outside area. In this context, the headset generates VR content. Alternatively, in some embodiments, one or both of the display elements are at least partially transparent, such that light from the local area may be combined with light from the one or more display elements to produce AR and/or MR content.

[0005] The eye tracking system is configured to track gaze locations of the user during an artificial reality experience. The eye tracking system may include, e.g., one or more cameras, a plurality of electrodes, some other system for tracking eye orientation, or some combination thereof. For example, in embodiments where the eye tracking system includes one or more cameras, the one or more cameras may be integrated into the headset so that they are inward facing and can capture images of the eye. And in embodiments, where the eye tracking system includes a plurality of electrodes, the electrodes may be integrated into the headset (e.g., temple, nose area, around the ear area, etc.) and/or one or more in-ear devices (e.g., to record eye movement). The signals from the plurality of electrodes may be used to generate an

electrooculogram (EOG) for determining eye position. In embodiments, where the IEDs and the headset have electrodes, the EOG artifacts may be removed (via a calibration) from electrooculogram (EOG) signals captured by the electrodes within the IEDs.

[0006] A speaker array provides audio content to the user. The speaker array may be integrated into the headset, the IEDs, or both. The speaker array is configured to provide spatialized audio content to the user in accordance with instructions from the controller. In some embodiments, loudspeakers are placed on the two sides of the headset to provide immersive audio contents for the user.

[0007] The controller controls components of the system. The controller spatializes audio content for presentation via the speaker array and sensory information from a head-tracking system (e.g., IMU, outside-in head-tracking or inside-out head-tracking system, etc.). The presented spatialized audio content may be congruent with visual content presented to the user (i.e., the spatialized sound is perceived to come from a target location where a visual object is presented and would appear to be generating the spatialized audio). In order to spatialize audio content, the controller may determine sound filters for the speaker array. The sound filters cause the audio content to be spatialized, such that the audio content appears to originate from a target location. The target location corresponds to a location where a visual object appears to be located (as presented via the display assembly) and generating sound. For example, the visual object may be a view of a virtual or real person speaking, and the target location is a mouth of the virtual or real person such that the spatialized audio appears to originate from a portion of the visual object, that if real, would generate the sound. The controller may use HRTFs and/or acoustic parameters to generate the sound filters. The acoustic parameters describe acoustic properties of the local area. The acoustic parameters may include, e.g., a reverberation time, a reverberation

level, a room impulse response, etc. In some embodiments, the controller calculates one or more of the acoustic parameters and/or retrieve them from a server. The controller instructs the speaker array to present the spatialized audio content.

[0008] The controller performs one or more realism checks while audio content is being presented to the user. In some embodiments, the controller performs a realism check where audio contents from different target 3D locations are presented to the user. While the contents are presented to the user, the user's audio perceptions are captured using electrode assemblies (the content of the captured EEG information will vary based on the presented 3D location of the sound). Therefore there is a closed-loop between the target (ground-truth) 3D location of the sound and also the measured electrophysiological signals captured using the electrode assembly. Once this realism check is established, the controller can use the parameters of the realism check to self-correct and adjust for the audio render pipeline to minimize the error. The controller may perform a realism check on a set schedule (e.g., every X minutes of operation). In other embodiments, the controller randomly performs a realism check. In some embodiments, the controller may randomly check the realism based on a previously-calibrated information on either regular cadence (e.g., every 1 minutes) or irregular basis (e.g., when contents are changed, the room acoustic information are changed, etc.). For a given realism check, the controller adjusts the spatialized audio content that is spatialized to the target position such that it is temporarily (e.g., 5 second) spatialized to a test location for a period of time and then reverts back to being spatialized at the target location. The test location is substantially different from the target location, e.g., 20° or more apart in azimuth. The change in spatialization audio causes an unexpected and sudden incongruity between where the visual object appears to be located and where the user is perceiving its sound to originate. For example, the virtual or real person may

appear to be speaking to the user from a target location (e.g., in front of user, 1.5 meters away)– and suddenly the voice of the virtual or real person appears to be originating from a location that is different (e.g., to the user's immediate right) from the target location. Note this is a temporary adjustment in spatialization, and it quickly reverts back to being spatialized to the target position. In cases where realism of the artificial reality experience is high, a gaze location of a user tends to fixate on the target location. And in these cases, a break in congruency causes an involuntary movement of the gaze location from the target location to the test position within a short time period (e.g., approximately the first 80-250 ms) immediately after a switch point. The switch point is a time when the system stops presenting audio spatialized to the target location and instead presents audio spatialized to the test position. Note that the timing of the eye movement is consistent with the interpretation that a cortical mechanism underlies the eye movement. The cortical mechanism may be accessible via EEG captured using electrodes on the IEDs, embedded within the headset itself, or both. In contrast, if the realism of a the artificial reality experience is relatively low, such a break in congruency tends not to generate the involuntary movement of gaze location within the initial period of time – but may generate voluntary movement of the eye toward the test location that occurs after the initial period of time.

[0009] When realism of the artificial reality experience is high, some users may choose to compensate for the involuntary eye movement by increasing their effort to maintaining gaze. The EOG tracks the user's intent to compensate.

[0010] The controller is configured to determine a level of realism of the artificial reality experience based in part on the tracked gaze locations, the EOG and/or EEG. The controller provides inputs (e.g., the tracked gaze locations before the switch point, the tracked gaze locations after the switch point, the target location, the test location, signals from the plurality of

electrodes, or some combination thereof) into a realism model which determines an associated level of realism. The controller may, e.g., use the determined level of realism to monitor performance of the system with regard to providing a realistic artificial reality experience to the user. Note that in some embodiments, the controller may re-calibrate the audio to improve realism if one or more realism checks are below a threshold level. For example, the controller may adjust one or more HRTFs to improve spatialization of audio content for the user.

[0011] Embodiments of the invention may include or be implemented in conjunction with an artificial reality system. Artificial reality is a form of reality that has been adjusted in some manner before presentation to a user, which may include, e.g., a virtual reality (VR), an augmented reality (AR), a mixed reality (MR), a hybrid reality, or some combination and/or derivatives thereof. Artificial reality content may include completely generated content or generated content combined with captured (e.g., real-world) content. The artificial reality content may include video, audio, haptic feedback, or some combination thereof, any of which may be presented in a single channel or in multiple channels (such as stereo video that produces a three-dimensional effect to the viewer). Additionally, in some embodiments, artificial reality may also be associated with applications, products, accessories, services, or some combination thereof, that are used to create content in an artificial reality and/or are otherwise used in an artificial reality. The artificial reality system that provides the artificial reality content may be implemented on various platforms, including a wearable device (e.g., headset) connected to a host computer system, a standalone wearable device (e.g., headset), a mobile device or computing system, or any other hardware platform capable of providing artificial reality content to one or more viewers.

Additional Configuration Information

[0012] The foregoing description of the embodiments has been presented for illustration; it is not intended to be exhaustive or to limit the patent rights to the precise forms disclosed. Persons skilled in the relevant art can appreciate that many modifications and variations are possible considering the above disclosure.

[0013] Some portions of this description describe the embodiments in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are commonly used by those skilled in the data processing arts to convey the substance of their work effectively to others skilled in the art. These operations, while described functionally, computationally, or logically, are understood to be implemented by computer programs or equivalent electrical circuits, microcode, or the like. Furthermore, it has also proven convenient at times, to refer to these arrangements of operations as modules, without loss of generality. The described operations and their associated modules may be embodied in software, firmware, hardware, or any combinations thereof.

[0014] Any of the steps, operations, or processes described herein may be performed or implemented with one or more hardware or software modules, alone or in combination with other devices. In one embodiment, a software module is implemented with a computer program product comprising a computer-readable medium containing computer program code, which can be executed by a computer processor for performing any or all the steps, operations, or processes described.

[0015] Embodiments may also relate to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, and/or it may comprise a general-purpose computing device selectively activated or reconfigured by a computer program

stored in the computer. Such a computer program may be stored in a non-transitory, tangible computer readable storage medium, or any type of media suitable for storing electronic instructions, which may be coupled to a computer system bus. Furthermore, any computing systems referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[0016] Embodiments may also relate to a product that is produced by a computing process described herein. Such a product may comprise information resulting from a computing process, where the information is stored on a non-transitory, tangible computer readable storage medium and may include any embodiment of a computer program product or other data combination described herein.

[0017] Finally, the language used in the specification has been principally selected for readability and instructional purposes, and it may not have been selected to delineate or circumscribe the patent rights. It is therefore intended that the scope of the patent rights be limited not by this detailed description, but rather by any claims that issue on an application based hereon. Accordingly, the disclosure of the embodiments is intended to be illustrative, but not limiting, of the scope of the patent rights, which is set forth in the following claims.

What is claimed is:

1. A method comprising:

tracking gaze locations of a user during an artificial reality experience;

adjusting spatialized audio content that is spatialized to a target position associated with a virtual object such that it is temporarily spatialized to a test location for a period of time and then reverts back to being spatialized at the target location, wherein the test location is different from the target location;

presenting the spatialized audio content, wherein the virtual object is concurrently displayed to the user at the target location, and over the period of time the audio content is spatialized to the test location and then reverts back to the target location; and

determining a level of realism of the artificial reality experience based in part on the tracked gaze locations within a portion of the period of time.