June 2022

Automatically Splitting and Routing Foreground and Background Audio to Different Output Devices for Immersive Surround Sound

D Shin

Automatically Splitting and Routing Foreground and Background Audio to Different Output Devices for Immersive Surround Sound

ABSTRACT

In the head-lock mode, headphones operate such that the directionality of sound sources within the audio is locked to the location of the headphones and not to locations within a physical scene. The head-lock mode can reduce the realism of the surround sound simulation due to failure to take into account the rotation or tiling of the user's head to adjust the directionality of the sound sources. This disclosure describes techniques to automatically provide a more physically accurate surround sound experience by splitting the foreground and background parts of an audio stream. The foreground audio is routed to the user's headphones in passthrough mode while the background audio is played via external speakers in the user's environment. The splitting can be achieved using suitably trained machine learning models or other techniques.

KEYWORDS

- Surround sound
- Immersive audio
- Foreground audio
- Background audio
- Directional audio
- Audio stream spitting

BACKGROUND

Immersive audio delivered via surround sound technology can enhance the audio experience of multimedia content, such as music, movies, videoconferences, etc. Such an experience can be provided even when a user is listening to the audio via headphones. However, in current implementations, the surround sound experience delivered via a user's headphones operates in a head-lock mode. Such operation can create the experience of the various sound sources that are part of the audio feed appearing to originate from different directions in relation to the location of the headphones. In other words, the directionality of the various sound sources within the audio feed is locked to the location of the user's headphones and not to locations within the physical scene.

The head-lock mode can reduce the realism of the surround sound simulation because a realistic surround sound experience must take into account the rotation or tilting of the user's head to make corresponding adjustments to the directionality of the sound sources in the physical world, such as people conversing, speakers playing music, etc. In other words, a realistic surround sound simulation needs to operate in a mode that locks the directionality of the sound based on positionality within the physical world.

Some headphones include motion sensors, such as an inertial measurement unit (IMU), that can be employed in conjunction with head-related transfer functions (HRTF) for dynamically adjusting the positionality of sound sources within the immersive audio experience. However, the IMUs measure relative movement and acceleration rather than absolute displacement, thus leading to drift in the output.

DESCRIPTION

This disclosure describes techniques to create a physically accurate surround sound experience for a user consuming immersive audio via headphones. To do so, speakers within the user's physical environment are leveraged in addition to headphones. For instance, such speakers can include the speaker of the television present in the room in which the user is located. Such a setup results in proximal real-life sounds being played closer to the user (via the user's headphones) while ambient sounds are delivered in the background (via an external speaker), thus providing a physically accurate and more realistic surround sound experience.

The setup described above is achieved by processing the immersive audio stream to extract foreground and background sounds within the audio stream. For instance, foreground audio can include human speech and other such sounds with high information content while background audio can consist of ambient sounds, environmental noise and other such sounds with low information content. The foreground audio is played via the user’s headphones while the background audio stream is routed to the speaker within the user’s physical surroundings. Further, the headphones are set to operate in passthrough mode to ensure that the user can hear the background stream being played through the external speaker.

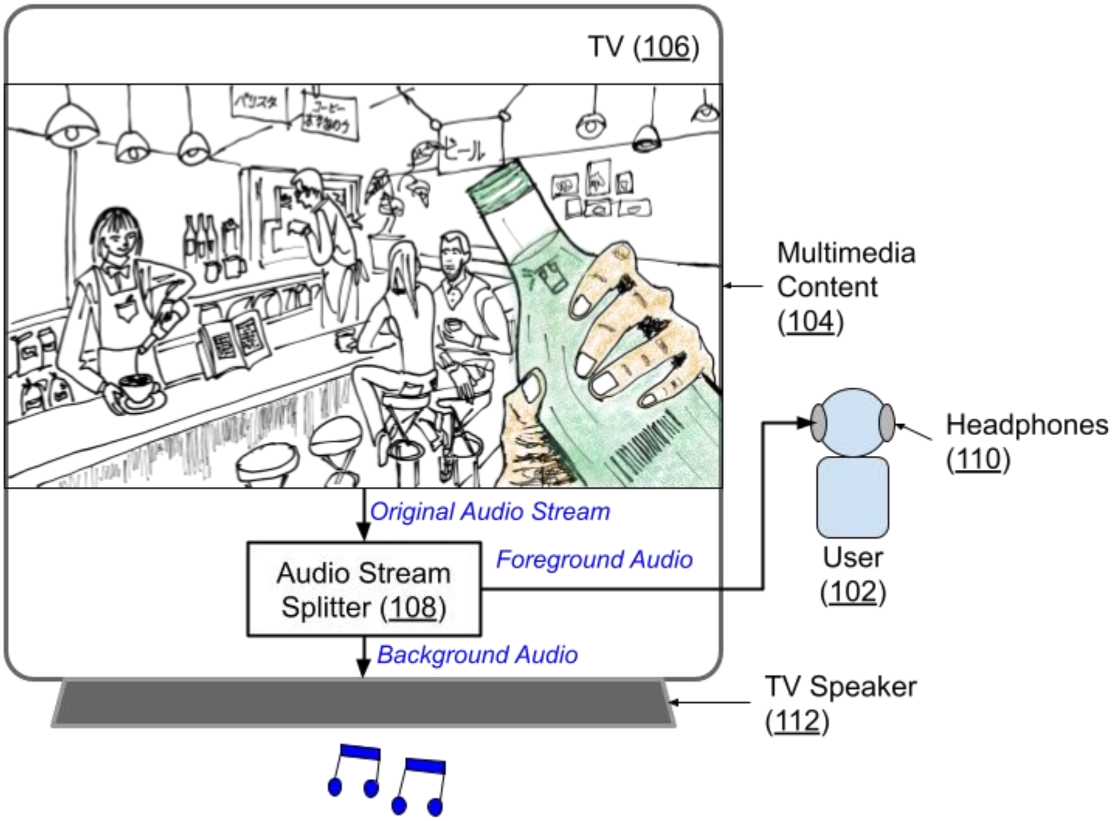

Fig. 1: Splitting and routing foreground and background audio within an audio stream

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user (102) is viewing multimedia content (104) on a TV (106). The original audio stream for the content is split by an audio stream splitter (108). The foreground audio of the people talking within the scene routed to the user's headphones (set to passthrough mode) while the background audio of the ambient music playing at the location within the scene is played via the external TV speaker (112). Splitting and routing of the foreground and background audio to different audio output devices can provide a more realistic surround sound experience.

Since the audio output devices used for the two separated parts of the original stream are located at different distances from the user, the playback of the two streams needs to be synchronized to avoid the audio stream from the farther device reaching the user later than the corresponding portions of the audio stream playing via the closer device. Such synchronization is achieved via a finite impulse response (FIR) all-pass filter applied to the portion of the audio stream played through the output device closer to the user. For instance, in the scenario depicted in Fig. 1, the delay filter is applied to the foreground audio played via the user's headphones to synchronize their output with that of the TV speaker that plays the background portion of the stream. The value of the delay can be determined by initial setup steps in which the user provides estimates of the distances of external audio sources from typical locations (e.g., couch) from which the user views multimedia content. Alternatively, or in addition, the appropriate delay interval can be derived dynamically at runtime.

Separation of foreground and background audio within a given sound stream can be achieved via a scoring model trained in a supervised manner using offline labeled data of single-channel sound snippets. Experienced annotators can mark the times within the snippets that contain foreground audio, such as people speaking while the remainder of the stream is considered to be labeled as background sounds. A sufficiently large amount of audio snippets labeled in such a manner are used for training the audio splitting model with a hyperparameter specifying a reasonable sliding window size, such as 0.1 second.

The window is slid over the labeled dataset with a stride, with consensus voting used to pick the most appropriate label (i.e., foreground or background) within a window. Repeating the process for all windows in all labeled recording sessions yields a matrix that stores triplets of: (1-second aggregate audio, 1-second foreground audio, 1-second background audio). Each aggregate audio snippet is then converted into a corresponding spectrogram to train a convolutional autoencoder (or other suitable model) that is trained to separate foreground and background audio components. The trained audio stream splitting model can be stored and utilized locally on any relevant user device, such as a smart TV, smartphone, tablet, etc., that can play multimedia content with surround sound.

The techniques described in this disclosure can support any devices and sound output sources capable of playing audio content. The techniques can be employed for any type of audio content such as music, videos, videoconferences, audio calls, etc. Implementation of the techniques can provide immersive audio playback locked to physical locations for directionality, thus enhancing the user experience (UX) of viewing multimedia content with surround sound.

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 headphones and available speakers, video content, 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 automatically provide a more physically accurate surround sound experience by splitting the foreground and background parts of an audio stream. The foreground audio is routed to the user's headphones in passthrough mode while the background audio is played via external speakers in the user's environment. The splitting can be achieved using suitably trained machine learning models or other techniques.