November 2022

ENABLING VIRTUAL ACOUSTIC BACKGROUND FOR VIDEO AND AUDIO CONFERENCING

Michelle Mao

Ivana Balic

Samir Ouelha

Pengfei Sun

ENABLING VIRTUAL ACOUSTIC BACKGROUND FOR VIDEO AND AUDIO CONFERENCING

AUTHORS:
Michelle Mao
Ivana Balic
Samir Ouelha
Pengfei Sun

ABSTRACT

Virtual background is an emerging feature of collaboration services such as conference calling applications and platforms. It enables users to choose a picture or video as their background to avoid distractions and protect privacy. In this article, we propose to enable virtual acoustic background for video/audio conferencing system. Such a feature can improve speech clarity and intelligibility for conference participants by making the collaboration more efficient and professional.

DETAILED DESCRIPTION

As an initial matter, it will be helpful to confirm the meaning of an element of nomenclature. The discussion below refers to conference call capabilities within an online communication and collaboration service. Such a service, which for simplicity of exposition may be referred to herein as a collaboration service, brings together different capabilities such as video conferencing, online meetings, screen sharing, webinars, Web conferencing, and calling.

The world is moving towards a hybrid work model. Within such a model, a positive conference call experience may be the most essential component. Many collaboration service conference calling platforms have enabled an interesting feature called virtual background. Such a feature allows a user to replace their real background with that of a picture that they like. A virtual background may help in a case where the real background is too distracting to show or when a user does not wish to show from where they are making a call. However, a similar feature – but for an acoustic background – is missing in collaboration service conference call platforms and applications.

To address such issues, techniques are presented herein that support a virtual acoustic background through which a user may experience an improved clarity and intelligibility of their speech during a conference call.

When users spend an extensive number of hours on video and audio calls with their colleagues, a decent acoustic background becomes extremely important for good collaboration. Such an acoustic background often comprises two aspects – distracting noises and reverberation. To fight against distracting noises, an individual may, whenever possible, make a call in a quiet room or turn on a noise-removal function in a conference call application if such a function is available. To deal with reverberation, suggestions from acoustic engineers may include using carpet rather than hard floors, placing a bookcase or a frame on the wall to either reduce the reflections or increase the absorption rate, simply moving closer a microphone, or employing a headset. Unfortunately, there are few software solutions or features in any collaboration services that address dereverberation issues during conference calls (such solutions or features having not shown evidence of positive results or existing as "black box" solutions).

Reverberation, which is created by sound waves being reflected from surfaces, is a common effect that is perceived in sound in everyday life. Reverberation is often characterized by three components – direct sound, early reflections, and late reverberations – as depicted in Figure 1, below.

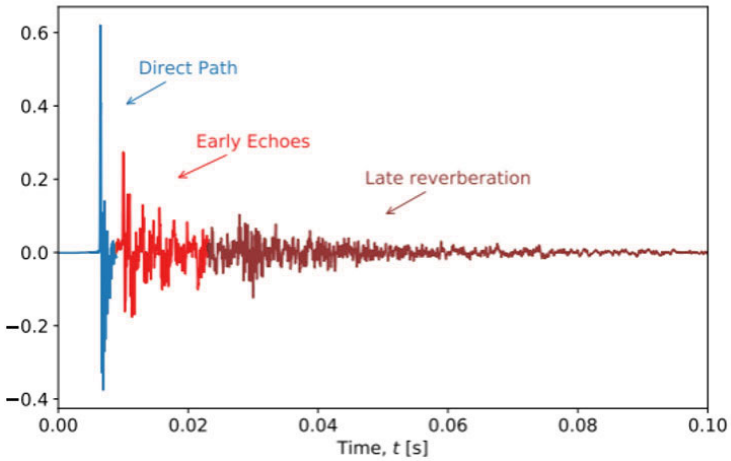

Figure 1: Exemplary Reverberation Components

In the exemplary reverberation components that are illustrated in Figure 1, the direct path is shown in blue, the early echoes are shown in red, and the late reverberation is shown in brown. The direct sound corresponds to the signal that travels directly from a source to a receiver along a straight line. Early reflections are defined as the sound waves that arrive at a receiver within the first 10 to 80 milliseconds (ms) after the direct sound.

The techniques presented herein support two different methods for removing an existing acoustic background and adding a new virtual acoustic background during conference calls and meetings.

A first method comprises two stages. Figure 2, below, presents a block diagram that depicts one possible arrangement of those stages that is possible according to the techniques presented herein.

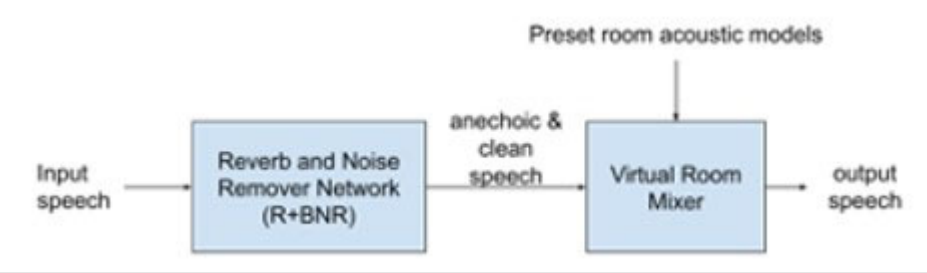

Figure 2: Block Diagram of Two-Stage Virtual Acoustic Background System

As shown in Figure 2, above, during the first stage any background noise and reverberations may be removed using neural networks. The output from the first stage will be anechoic and clean speech. In the second stage, a chosen room impulse response (RIR) may be applied to the anechoic and clean speech. The resulting output speech will sound more intelligible, richer, and more realistic compared to its anechoic version. There are a number of ways to select an appropriate RIR, and some of those ways will be described later in the instant narrative.

In a more general sense, during the second stage noise may be added in a controlled fashion before applying an RIR. Adding noise may be useful in a case where it is desirable to reproduce some realistic background such as, for example, a church, a concert hall, a hospital, etc.

A second method performs the removal of the existing background and the replacement with a new background all in one step. A neural network may be trained to

remove the noise and reverberations and then apply a chosen RIR, and possibly a noise, to the output. An advantage of this method is that it is simpler, and it may be easier for training a neural network.

Finally, to set the virtual acoustic background with additional parameters an equalizer may be added. Such an equalizer may encompass different preset modes (including, for example, a "voice boost" mode). Additionally, a user may manually set an equalizer by frequency bands (with a limited number of bins).

Aspects of the techniques presented herein encompass, among other things, the training of a neural network.

To address the dereverberation problem through a machine learning (ML) and deep learning (DL) method, according to the techniques presented herein, emulated reverberated speech (which may be developed by applying an RIR to clean anechoic speech) may be used to train a dereverberation model. To obtain a large scale of diversified and realistic RIRs, aspects of the presented techniques employ a combination of real-world RIR datasets (which can be found on the Internet) and simulated RIR data that may be obtained through different types of model, like exponential-type RIR, stochastic RIR or Image-Source Method (ISM), as described in published literature dealing with this subject.

As illustrated above, the techniques presented herein support two different methods for removing an existing acoustic background and adding a new virtual acoustic background during conference calls and meetings. A neural network training procedure for the two different methods is described below.

For the first method, the input training data comprises an anechoic clean speech, to which different types of noises are added and an RIR is applied. The noise types and the RIR should be diverse to simulate realistic acoustic backgrounds. The target for the first method is the anechoic clean speech. After the clean and anechoic speech signal is obtained, it is possible to create a virtual room experience by applying an RIR that is chosen from some preset acoustic models.

For the second method, the target is the anechoic clean speech that is convolved with a "golden" RIR. Further, a chosen noise may be added before the convolution with the RIR.

It is important to note that removing reverberation and background noise is a fundamental step. If an RIR were to be applied to noisy or reverberated speech, the intelligibility would be significantly degraded because of the “boosted” reverberation and noise.

As noted previously, aspects of the techniques presented herein support the selection of an RIR and an optional background noise. To obtain a preset RIR, subjective tests may be designed to identify a number of “optimal room settings” that maximize a human opinion score in terms of speech quality or intelligibility. Additionally, some number of “fun room settings”" (such as a church, a concert hall, a forest, a beach, etc.) may be included as additional choices. Such settings do not aid in the pursuit of better speech quality or intelligibility, but they make the technology of the techniques presented herein more attractive for non-work-related calls.

As described above, further aspects of the techniques presented herein support the deployment of a new virtual acoustic background. A user may choose between a set of predefined acoustic backgrounds, or the collaboration service may make a default choice. When a user chooses from the existing acoustic backgrounds, the procedure comprises a series of steps. Under the first step, a user makes a short recording of herself or himself. Then, during the second step, the software applies an existing virtual background to the recorded signals. Finally, during a third step, the user listens to the processed signals and then selects the one that he/she likes the most.

Optionally, the available acoustic virtual backgrounds may have descriptive names and the user may select one of those background without first checking how it sounds with her or his own voice.

According to aspects of the techniques presented herein, in a single-talker scenario a single RIR may be applied to create a virtual acoustic background. In a multi-talker scenario, different RIRs may be applied to different speakers in order to better reproduce their spatial correlations in the scene.

Use of the techniques presented herein offers several advantages. As a first advantage, the presented techniques may improve the remote work experience better than an in-person interaction through the addition of virtual acoustic background. Since it is not possible for an individual to change the acoustic model of an existing room or space, they

can simply enable the technology that is supported by aspects of the presented techniques. As a second advantage, the presented techniques make non-professional calls more interesting by enabling users to select an acoustic setting that they would like to be virtually.

In summary, techniques have been presented herein that support a virtual acoustic background through which a user may experience an improved clarity and intelligibility of their speech during a conference call. Aspects of the presented techniques support the removal of background noise and reverberations by a neural network, the training of such a neural network, the application of a selected RIR, etc. Use of the presented techniques make a remote work experience better than an in-person interaction through the addition of a virtual acoustic background and make non-professional calls more interesting by enabling users to select a virtual acoustic setting that they would like.