August 24, 2016

"AUTOMATIC SELECTION OF BINAURAL OR MONAURAL AUDIO SOURCES IN VIDEO CONFERENCING SYSTEMS"

Simon Smith

AUTOMATIC SELECTION OF BINAURAL OR MONAURAL AUDIO SOURCES IN VIDEO CONFERENCING SYSTEMS

ABSTRACT

Disclosed herein is a video conferencing system and method for collecting binaural and monaural feed simultaneously from the endpoints and selecting the best feed to provide each endpoint based on the capabilities of the endpoint. Each endpoint is tagged to indicate its receiving capability. When a call is started, the audio sources from the endpoints are collected simultaneously. An endpoint having a lone participant is provided with binaural feed while the endpoint having multiple participants is provided with monaural feed. The disclosed system and method provides a lone participant a more immersive audio experience and also allows them to better discriminate between different people in a room when many are talking at the same time.

BACKGROUND

Video conferences consist of joining multiple endpoints together. Traditionally, video conferences comprised two rooms, but recently this has changed. The conferences may include many endpoints in each call.The conference rooms at each endpoint may contain either multiple people or only one participant. The endpoints are usually configured to send and receive either binaural or monaural audio feeds. Also, the endpoints may consist of a mixture of endpoint types where there are multiple participants at one end and only a single participant at the other end.

In cases where there is a lone participant (who may be wearing headphones), the most immersive audio experience would involve receiving a stereo audio feed from a virtual listener feed from larger spaces capturing stereo audio. Conversely, it is difficult to recreate a stereo

image for all participants in a meeting room with multiple participants and so the meeting room would best be served by a mono audio feed.

The stereo feed should be an accurate capture of the audio in the room as if the listener were present in the room. The inter-aural timing of the feed should match that of a typical human listener. The feed could be a capture of a pair of microphones separated approximately ear distance apart or be captured through a soundfield microphone and processed through a head related transfer function. For the mono feed, the best collection is via microphones distributed across the length of the table. The stereo feed allows a lone participant to feel more immersed in the video conference and allows them to better discriminate between people in a room when many are talking at the same time.

There currently exists no system to collect both mono and stereo audio or correctly feed it to participants based on their capability.

DESCRIPTION

This disclosure provides a system and method for video conferencing endpoints to collect stereo and mono audio feeds simultaneously and then distributing the appropriate feed to each endpoint, thereby delivering the best audio experience for each specific type of endpoint. The system as shown in FIG. 1 includes a mixture of collection devices in the conference rooms, with each endpoint system capable of transmitting 3 channels of audio (L, R and Mono). A virtual listener system in the room places the microphones at a location that mimics a person in the space. The mono audio mixes are arranged and processed so as to provide a normalized level mono mix to the participants.

The best method for collecting the stereo feed at the endpoints is to place a pair of microphones approximately ear distance apart to mimic a participant at the table. For the mono feed, the best collection method is by placing microphones distributed across the length of the table.

As shown in FIG. 1, the system comprises video conferencing endpoints, for example, two large conference rooms and a lone endpoint/participant. The call management system receives the binaural signals 1-LR-OUT and 3-LR-OUT from the large conference rooms and mixes them as binaural stereo signals 2-LR-IN to be provided to the endpoint having the lone participant. Also, the system mixes the mono outputs 1-MONO-OUT from the first endpoint, i.e. the first conference room and 2-MONO-OUT from the second endpoint, i.e. the endpoint having a lone participant and feeds the monaural audio signal 3-MONO-IN to the second conference room. Likewise, the system mixes the monaural signals 2-MONO-OUT and 3-MONO-OUT from the lone participant and the second conference room to feed the mono audio signal 1-MONO-IN to the first conference room. The processes of collecting the audio feeds, mixing and feeding the appropriate feeds to the participants occur in a simultaneous manner.

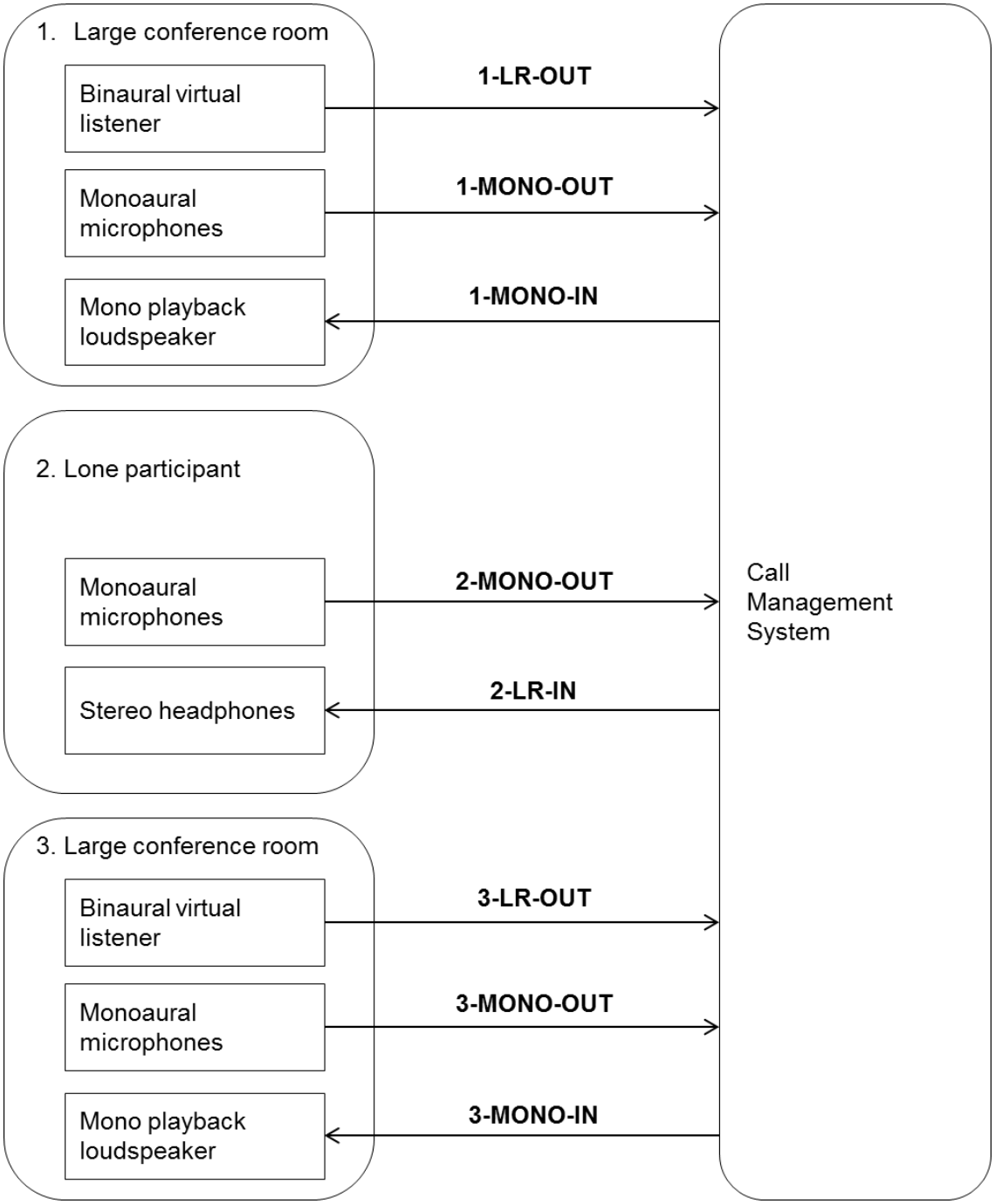

FIG. 1: System for collecting stereo and mono audio feeds and feeding it appropriately to the endpoint participants based on their capability

The method for automatically selecting binaural or monaural audio sources in video conferencing systems is illustrated in FIG. 2. In step A, each endpoint in the video conferencing

system is tagged. This will indicate whether a receiving endpoint would benefit from a stereo signal or a mono signal. In step B, when the call is initiated, the audio sources from the endpoints are collected simultaneously. In step C, the system determines the type of audio to provide to a particular endpoint, and in step D, the appropriate binaural or monaural signal is fed to the endpoint(s), i.e. those having a single participant are given a binaural signal while the endpoint(s) having multiple participants is fed with a monaural signal.

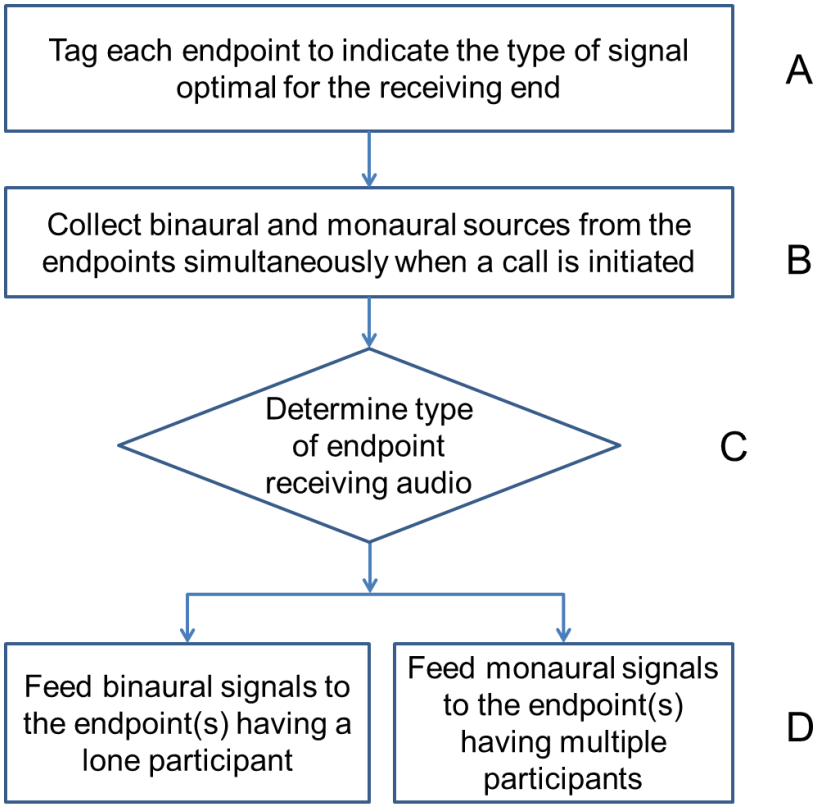

FIG. 2: Method for automatically selecting binaural or monaural audio sources in video conferencing systems

Thus, the disclosed system and method provides a more immersive audio experience and better ability to understand conversations with less impact on call bandwidth or infrastructure. Also the disclosure provides the possibility of placing each lone participant in different virtual

acoustic locations and enables them to more easily discriminate between different people's voices when multiple people are talking at the same time.