April 2020

Spatially Separating Participant Audio in a Conference Call

Jakob Zwiener

Marcos Calvo Lance

Jakub Kriz

Spatially Separating Participant Audio in a Conference Call

ABSTRACT

The audio signal transmitted by the far end of a conference call or video conference is typically received at the near end as a composite signal that is the sum of the audio signals of the far-end participants. Far-end participants are therefore not spatially separable at the near end. This prevents near-end participants from using the natural focusing abilities of the brain (cocktail party effect) to focus on the speech of particular far-end participants. This disclosure describes techniques, e.g., per-microphone audio channels, speech diarization, etc., that distinguish far-end participants such that their audio is spatially separated at the near end.

KEYWORDS

- Speech localization
- Sound spatialization
- Sound-source mapping
- Conference call
- Video conferencing
- Speech diarization
- Cocktail party effect

BACKGROUND

The audio signal transmitted from a far end of a conference call or video conference is typically received at the near end as a composite signal that is the sum of the audio signals of the far-end participants. Far-end participants are therefore not spatially separable at the near end. This prevents near-end participants from using the natural focusing abilities of the brain (cocktail party effect) to focus on particular far-end participants. Sometimes the constituent audio signals of far-end participants are so thoroughly mixed, e.g., when multiple participants speak simultaneously, that the resulting audio signal becomes unintelligible at the near end.

Many existing video or audio conferencing setups feature multiple microphones and speakers. These are currently used to improve the distance to the participant, e.g., the signal-to-noise-ratio of the participants' speech signals.

DESCRIPTION

This disclosure describes techniques to distinguish the far-end participants of a conference call such that their speech is spatially separated at the near end.

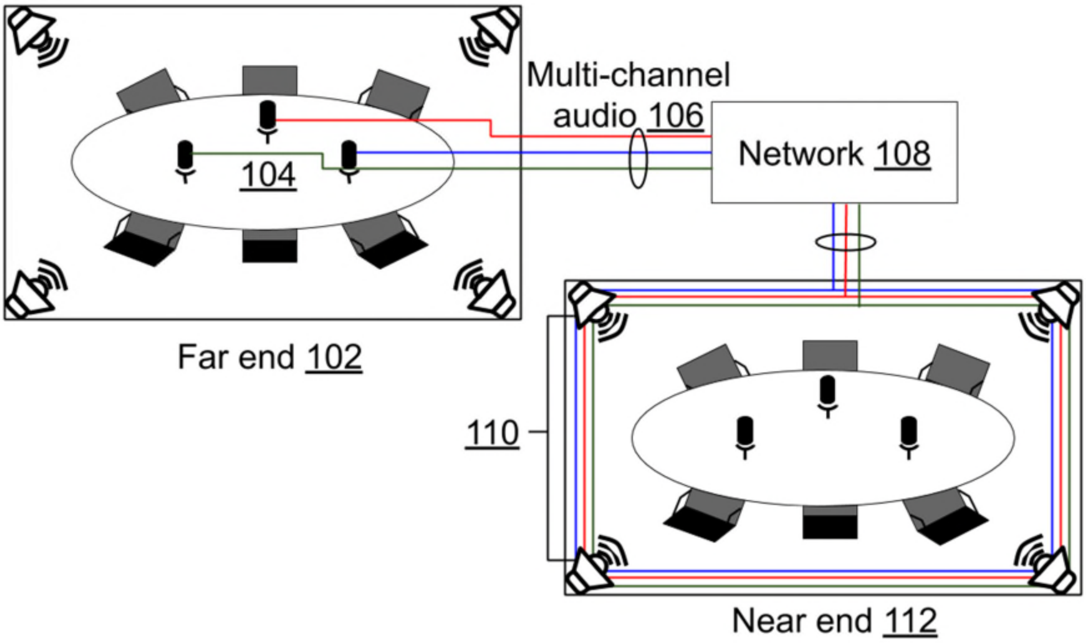

Fig. 1: Multi-channel audio for spatial separation of far-end participants

Fig. 1 illustrates an example technique for spatial separation of far-end participants. Signals from multiple microphones (104) at a far end (102) of an audio or video conference are captured and transmitted in the form of multi-channel audio (106) over a network (108) to the near end (112). The audio is reproduced by speakers (110) at the near end in a manner that enables near-end participants to spatially localize (directionally focus on) particular far-end participants (cocktail party effect).

For flexibility in the setup of microphones and speakers at the near-end and far-end conferencing rooms, room geometry is determined before the call, e.g., using a daily automatic test. This can be done, e.g., by creating impulses from the speakers and measuring latency through the microphones. The room geometry, e.g., distances and positions of the microphones, is used to generate the multi-channel audio that is transmitted to the near end and reproduced there using speakers.

In calls with more than two conferencing rooms, the different rooms are made spatially distinct in the receiving rooms by reconstructing the audio such that it appears to come from different directions. This can be done in conjunction with displaying a video of the far-end room at the side of the receiving room that is reproducing the corresponding audio.

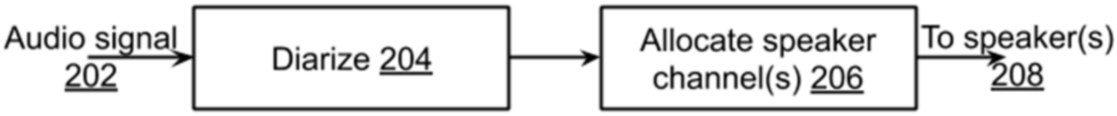

Fig. 2: Spatial separation of far-end participants using diarization

Fig. 2 illustrates another example technique for spatial separation of far-end participants, which is based on diarization. Speech diarization is a technique that segments an audio signal based on the identity of the speakers who appear in such audio. An audio signal can be diarized using machine learning or other techniques. The audio signal (202) that is a composite of the far-end participants of a conference call is diarized (204). A diarized segment is allocated (206) to one or more speakers (208) at the near end such that the participant on that segment is spatially localized.

With participant permission, the techniques described herein can also be used to identify individual far-end participants. This data can be used to split the audio into multiple channels that can be visualized and made controllable by near-end participants. This technique is especially useful for single users with headphones that don't have the facilities of a well-equipped conference room. Machine learning (or other techniques) can be used to enhance the quality of the audio on a per-participant basis. For example, the audio signal of an identified, soft-spoken participant can be increased relative to other participants.

CONCLUSION

This disclosure describes techniques, e.g., per-microphone audio channels, speech diarization, etc., that distinguish far-end participants such that their audio is spatially separated at the near end.