July 2023

Recovering Clean Speech from Noisy Audio Input via Ambient Speech Characterization

D Shin

Recovering Clean Speech from Noisy Audio Input via Ambient Speech Characterization

ABSTRACT

Video and audio conferencing hardware and software can implement noise filters to remove non-speech background noise captured when a participant is in a noisy environment. However, while noise filters can help remove non-speech background noise, the audio signal can still contain unwanted background speech from parties co-located with the speaker. The quality of a speaker's audio can additionally be impacted by hardware and/or software issues which noise filters cannot remove. This disclosure describes the use of generative voice models to clean up degraded audio of a user's speech. The recovery of the user speech is based on dynamically updated ambient characterization of the user's speech. With user permission, speaker embeddings are obtained by automatically segmenting conversations to identify the parts that involve a particular user's speech, without requiring the user to engage in a long calibration session. The noisy acoustic signal containing the user's speech and the rolling average of speaker embeddings characterizing the user's typical speech are input to a suitable vocoder-type neural network to obtain clean audio of the speaker's original speech.

KEYWORDS

- Speaker embedding
- Speech embedding
- Speech characteristics
- Background noise
- d-vector
- Short-Time Fourier Transform (STFT)
- Ambient speech
- Speech recovery
- Noise removal
- Acoustic segmentation
- Voice recovery
- Generative AI
- Generative model
- Vocoder

BACKGROUND

Audio and video conferences via smartphones, laptops, and other devices are common. In many cases, one or more users within an audio conference are in environments in which their speech is impacted by various external sounds, such as background music, speech of other people in the vicinity, traffic, etc. Some conferencing applications include noise filters that enhance speech within the audio signal while suppressing non-speech sounds. While such filters can help remove non-speech background noise, the audio signal can still contain unwanted background speech from parties that are co-located with the speaker. The quality of audio can additionally be impacted by hardware and/or software issues, such as defective microphone, lossy client-side compression, limited network bandwidth, etc. Filters that suppress background noise cannot mitigate such problems that degrade the quality of the audio that is provided to others in the conference.

Audio and video conferences are hosted by a server that appropriately relays the audio signals of various parties in the conference to each other, with a conferencing client application (or browser) on each user's device obtaining the audio signal for the respective user. The issues that degrade the acoustics of captured speech occur when detecting speech via the device microphone and/or relaying the audio from the client application to the conferencing server.

DESCRIPTION

This disclosure describes the use of generative voice models to clean up degraded audio of a user's speech in a conferencing application. With user permission, the clean user speech, after removal of environmental noise, can be recovered based on dynamically updated ambient characterization of the speech. Speaker embeddings are obtained by automatically segmenting conversations to identify the parts that involve a particular user's speech. The seamless nature of

the ambient operation avoids the need for users to engage in a long calibration session. The operation can be performed on the conferencing server since it is the last point in the conferencing path where client-side issues related to audio capture can impact the acoustic quality of a user's speech.

A d-vector corresponding roughly to the desired speaker embeddings for a user can be learned from a user's ambient speech, obtained with the user's permission. If the user permits, activation of the microphone on the client side of the conferencing application can be used as a natural trigger for acoustic segmentation during the learning phase. Subsequently, a sound detection model on the client and/or server side of the conferencing application can indicate whether the user is speaking during a given time window. For instance, the speech detection model for a given user can output a 0 or 1 to indicate whether the user is speaking during a time interval. The speech detection model can be on the client or server side. When the model computations are performed on the client, the results are pooled on the server.

Audio input captured during the time intervals when the output of the sound segmentation model indicates that the user is speaking is processed via Short-Time Fourier Transform (STFT) and a Convolutional Neural Network (CNN) for the speaker embeddings. The processing can generate a single embedding vector characterizing the user's speech. The embedding vector for the user's speech can be used to maintain a continually updated rolling average across the ambiently computed embedding vectors. The rolling average can serve to update speaker embeddings in real time by combining information regarding the user's most recent speech with relevant historic information about the user's speech characteristics. As a result, the characteristics of the user's speech that are learned can continually improve over time as the user speaks naturally.

The characteristics of a user's typical speech patterns within the rolling average of the user’s speech embeddings can be utilized to recover the user’s speech from the noisy acoustic signal captured by the conferencing client application. The noisy acoustic signal containing the user's speech and the rolling average of speaker embeddings characterizing the user’s typical speech can be input with permission to a suitable vocoder-type neural network, such as WaveRNN [1]. In contrast to typical signal denoising techniques that operate only on noisy input, the recovery of user speech utilizes additional input characterizing the user’s typical speech. The output of the neural network can provide audio that recovers the user’s speech from the noisy acoustic input by eliminating all forms of noise, such as background chatter, ambient sounds, glitches due hardware/software issues, audio-compression limitations, bandwidth constraints, etc. The recovered speech has accurate semantics and intonation as intended in the user’s original speech unaffected by noise.

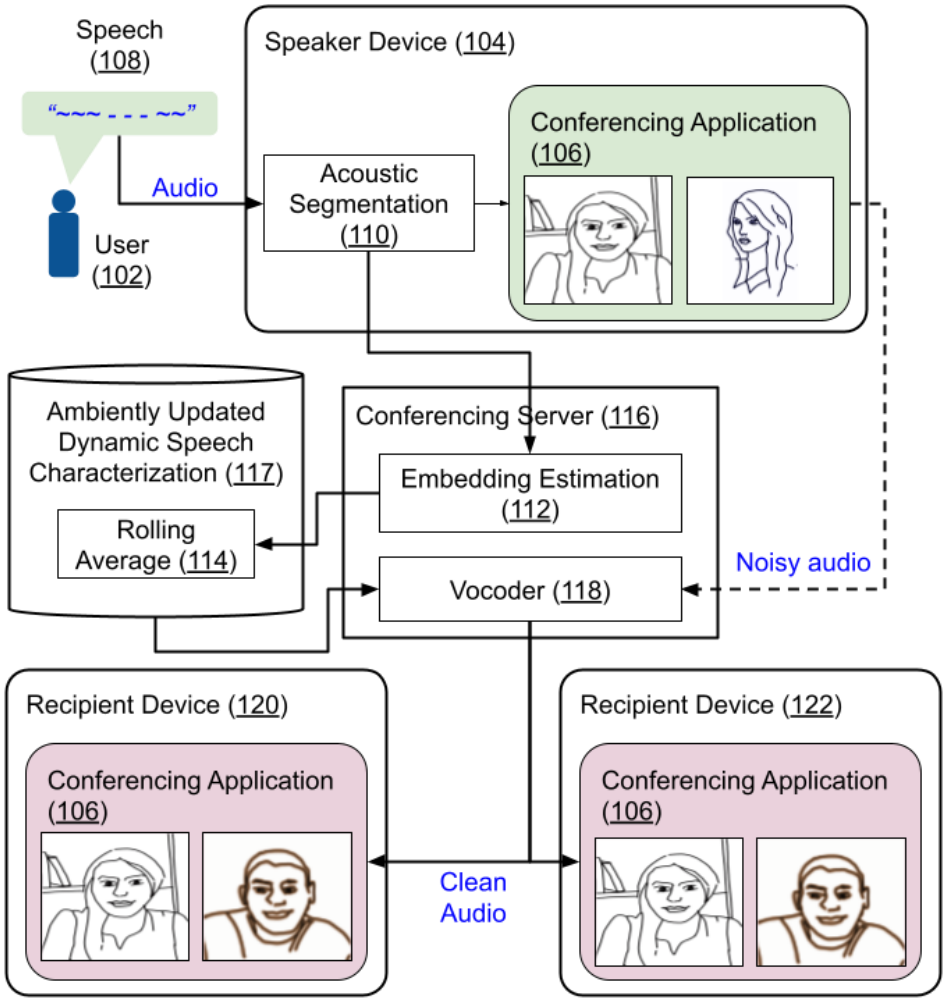

Fig. 1: Recovering original speech by noise removal via ambient speech characterization

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user (102) is speaking while participating in an audio/video conference via a device (104). The conference has two other participants, with respective devices (120 and 122). Acoustic segmentation (110) performed on the user's device that indicates the time intervals during which the user is speaking. The acoustic segmentation is used to estimate (112) the embeddings for the user's speech. If the speaker permits, a rolling average (114) of the estimated embeddings (114) is computed and maintained as a dynamically updated characterization of the speaker's typical speech (117). The noisy audio capture of the user's speech is relayed to the

conferencing server (116) by the client conferencing application (106) on the speaker's device. The noisy acoustic signal coupled with the characterization of the speaker's speech is input to a vocoder (118) to recover clean audio of the speaker's original speech without the noise. The recipient devices are provided with clean audio.

The techniques described herein can be implemented with user permission to support audio conferencing within any applications, including conference calls, video conferences, online meetings, messaging services, etc. The acoustic segmentation, embedding estimation, and dynamic speech characterization can be performed by the conferencing application on the client device and/or on the conferencing server, as appropriate. Implementation of the techniques described in this disclosure can seamlessly and dynamically enhance the clarity and quality of the audio of user speech within online conferences regardless of external environments in which users are located. The enhancement in audio quality of the speech of conference participants can significantly improve the user experience (UX) of online conferences.

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 speech characteristics, social network, social actions or activities, profession, 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 the use of generative voice models to clean up degraded audio of a user's speech. The recovery of the user speech is based on dynamically updated ambient characterization of the user's speech. With user permission, speaker embeddings are obtained by automatically segmenting conversations to identify the parts that involve a particular user's speech, without requiring the user to engage in a long calibration session. The noisy acoustic signal containing the user's speech and the rolling average of speaker embeddings characterizing the user's typical speech are input to a suitable vocoder-type neural network to obtain clean audio of the speaker's original speech.

REFERENCES

1. Kalchbrenner, Nal, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. "Efficient neural audio synthesis." In *International Conference on Machine Learning*, pp. 2410-2419. PMLR, 2018.