January 30, 2019

Enhanced call quality using user-specific voiceprint model

ChiLin Kuo

HsinHu Wang

SSU-TING TSAI

Chin-Kuo Huang

Enhanced call quality using user-specific voiceprint model

ABSTRACT

The audio quality of a call conducted using a user device such as a phone can be unsatisfactory due to a variety of reasons. Such reasons include noisy surroundings at the location from which the user conducts a call, position of the phone near the user’s face, users that speak with a soft voice, presence of far-end echoes, etc. This disclosure describes techniques that use a user-specific voiceprint to home into the user’s speech and cut out surrounding disturbances from a voice call. The techniques are implemented with user permission to generate and use the voiceprint.

KEYWORDS

- phone call
- call quality
- audio quality
- voiceprint
- voice features
- voice enhancement
- signal-to-noise ratio (SNR)
- noise suppression
- echo cancellation
- automatic equalization
- dynamic range compression

BACKGROUND

The audio quality of a call conducted using a user device such as a phone can be unsatisfactory due to a variety of reasons. Such reasons include noisy surroundings at the location from which the user conducts a call, position of the phone near the user's face, users that speak with a soft voice, presence of far-end echoes, etc.

DESCRIPTION

The voiceprint of a person is a set of features unique to the voice of that person. The voiceprint of a person can be learned and recognized by a machine learning model, known as a voiceprint model. The voiceprint model can be, e.g., a neural network. The techniques of this disclosure use the voiceprint to home into a user's speech and cut out surrounding disturbances.

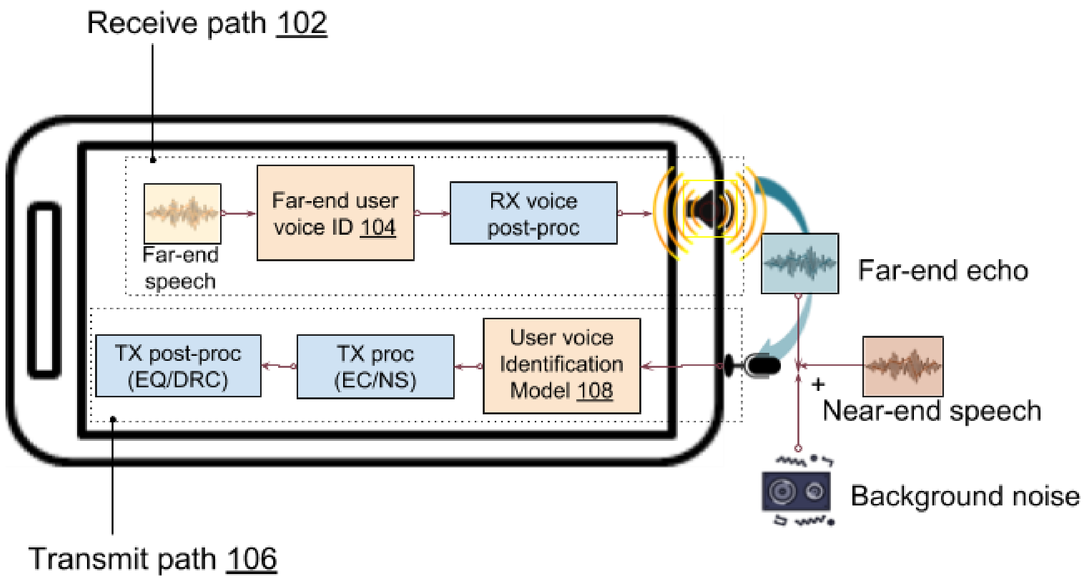

Fig. 1: Voiceprint model to enhance call quality

Fig. 1 illustrates the use of a voiceprint model to enhance call quality, per techniques of this disclosure. A voiceprint model (108) is inserted along the transmit path (106). In addition, a voiceprint model (104) can be inserted within the receive path (102). As shown, the transmitted signal includes not only the near-end speech, but also disturbances such as background noise and far-end echo. If the voiceprint model is inserted along the transmit path, e.g., before the echo-cancellation (EC) and noise-suppression (NS) blocks, it actively adapts on the user's speech signal during a call. If the devices at both ends of the call have voiceprint-based transmit path enhancement, then both ends experience improved voice quality. If the device at one end of the call has voiceprint-based transmit-path enhancement, then the far-end counterpart of the device experiences improved voice quality.

If the voiceprint model is inserted within the receive path, the voiceprint model passively utilizes voiceprint information at the beginning of receive voice processing. The voiceprint information is derived, for example, from databases located in the cloud. Voiceprint information is generated and used with user permission. If the device at one end of the call has both transmit and receive voiceprint enhancement, then voice quality at both that device and its far-end counterpart improve.

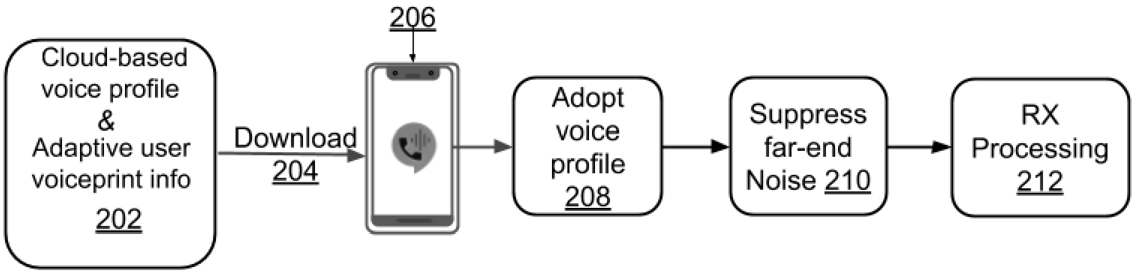
206
Cloud-based voice profile & Adaptive user voiceprint info 202
Download 204
Adopt voice profile 208
Suppress far-end Noise 210
RX Processing 212

Fig. 2: Receive path voiceprint processing

Fig. 2 illustrates an example of receive-path voiceprint processing. A cloud-based voice profile (202), which serves as an adaptive voiceprint, is downloaded (204) to a device (206) that participates in a call. The downloaded voiceprint is utilized as the far-end voice profile (208). This far-end voice profile is used within the receive path to suppress far-end noise (210). Following voiceprint-based far-end noise suppression, further receive processing is performed

(212).

Transmit-path voiceprint processing

Voiceprint processing can be included along the transmit path in different ways, such as, for example:

- **Prior to echo-cancellation/ noise suppression (EC/NS):** If the voiceprint model identifies user-specific voiceprint prior to EC/NS, then both linear and non-linear EC/NS can be performed with a pure user signal.
- **After linear EC/NS processing:** In this case, the voiceprint model identifies and enhances speech based on both non user-specific and user-specific signals.

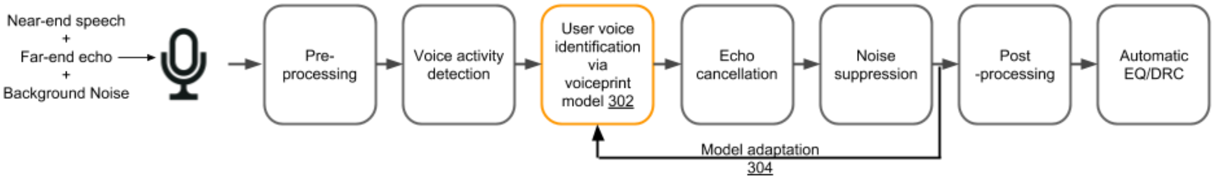

Fig. 3: Transmit-path voiceprint processing prior to EC/NS

Fig. 3 illustrates transmit-path voiceprint processing prior to EC/NS. The transmit path accepts as input a signal comprising near-end speech, far-end echoes, and background noise. As illustrated, the voiceprint model (302) identifies and enhances the user’s voice prior to echo cancellation and noise suppression. The voiceprint model is adapted (304) using the post-EC/NS signal, e.g., a signal that is relatively free of far-end echoes and background noise.

Traditional EC/NS can distort near-end speech, especially if done aggressively. The use of the voiceprint model prior to EC/NS enables identification of user voice segments accurately and preserves those segments during the subsequent processing. After EC/NS, user voice is further enhanced by automatic equalization and dynamic range compression (EQ/DRC) to perform perception recovery.

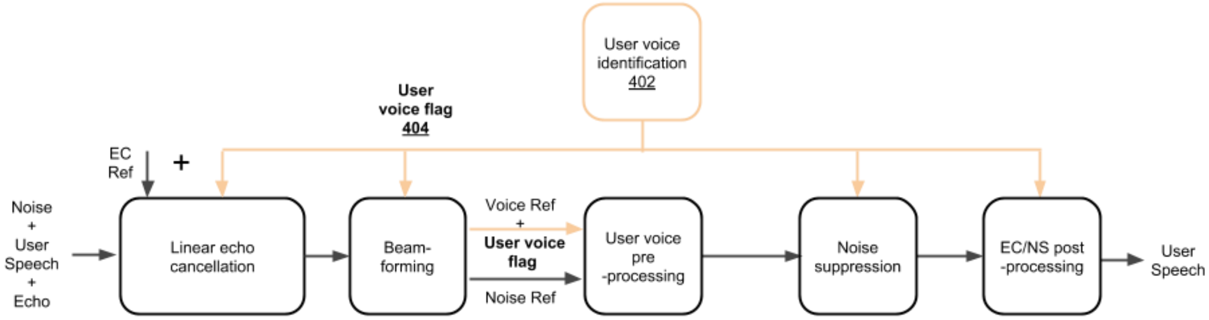

Fig. 4: Precise echo cancellation and noise processing using transmit voiceprint processing

As illustrated in Fig. 4, transmit-path voiceprint processing enables precise echo cancellation and noise suppression. The voiceprint model identifies the user's voice (402), after which it generates a voice flag for every segment where the user's voice is detected. By identifying segments of active speech, the voice flag enables optimal EC/NS processing during silent segments, and thereby improves the quality of EC/NS. As illustrated, the voice flag feeds into sub-blocks of the EC/NS path, e.g., linear echo cancelation, beamforming, voice pre-processing, noise suppression, and EC/NS post-processing. The voice flag is particularly helpful in the post-processing stage, which has no processing reference (unlike linear echo or noise cancellation) and is hence susceptible to introducing degradation.

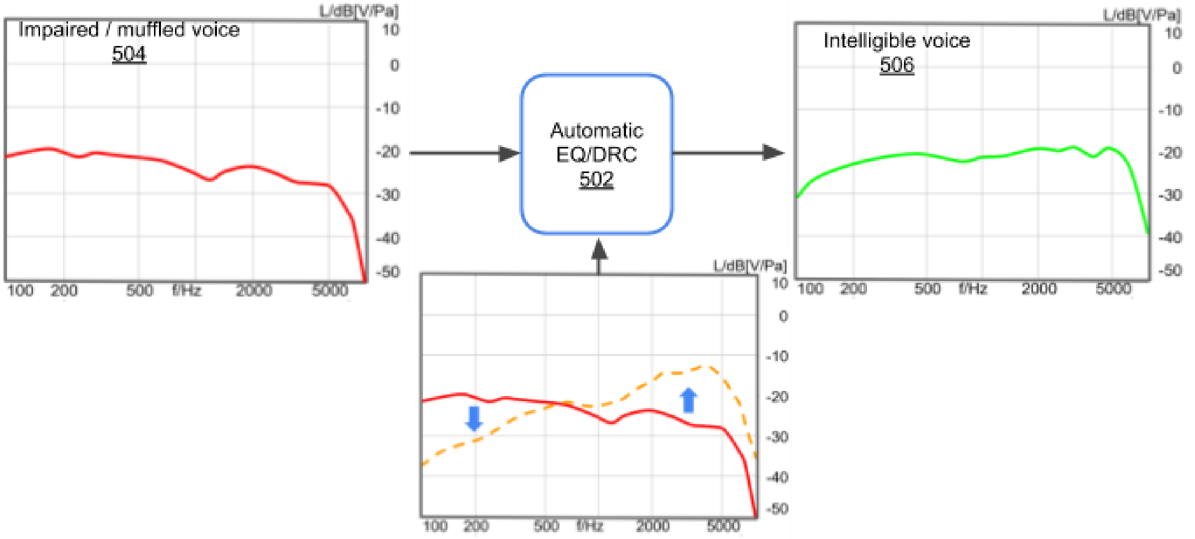

Fig. 5: Equalization and dynamic range compression using voiceprint model

As illustrated in Fig. 5, transmit-path voiceprint processing also helps in automatic equalization (EQ) and dynamic range compression (DRC). After EC/NS processing, EQ/DRC is applied (502) to recover impaired speech and muffled voice (504) based on user voiceprint. This boosts the speech segment level to provide more intelligible voice (506) without increasing the noise floor.

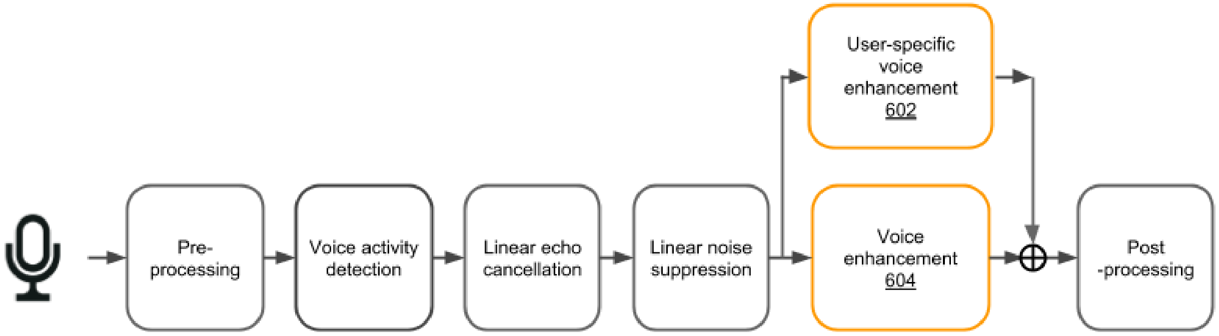

Fig. 6: Transmit-path voiceprint processing after EC/NS

Fig. 6 illustrates transmit-path voiceprint processing after EC/NS. A voiceprint model trained over a relatively large corpus of speech sources identifies speech. At the same time, a user-specific voice enhancement module (602) identifies the user’s voice using a model trained by the user’s own voiceprint. This user-specific enhanced speech is added to the output of a voice enhancement block (604) that enhances the EC/NS output.

Voiceprint model training

Per techniques of this disclosure, the voiceprint model can be trained in an offline manner or in an online manner.

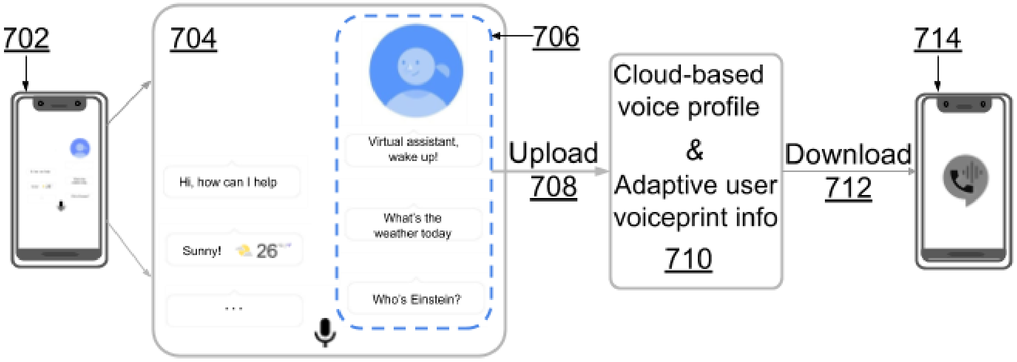

Fig. 7: Training the voiceprint model in an offline manner

Fig. 7 illustrates the training of the voiceprint model in an offline manner. A user orally interacts with a virtual assistant provided on a mobile device (702) such as a smartphone or other device. The utterances or commands of the user (706) that occur during the conversation (704) with the virtual assistant are uploaded (708) with user permission to develop a voice profile (710) of the user. This voice profile is a set of features that includes characteristics of the user's speech and serves as a voice ID or voiceprint of the user. The voiceprint adapts over voice data and becomes more accurate as more voice data is collected. During a call made by the user using a mobile device (714), the voiceprint is downloaded (e.g., from cloud-based voiceprint storage) and utilized (712) to enhance voice quality.

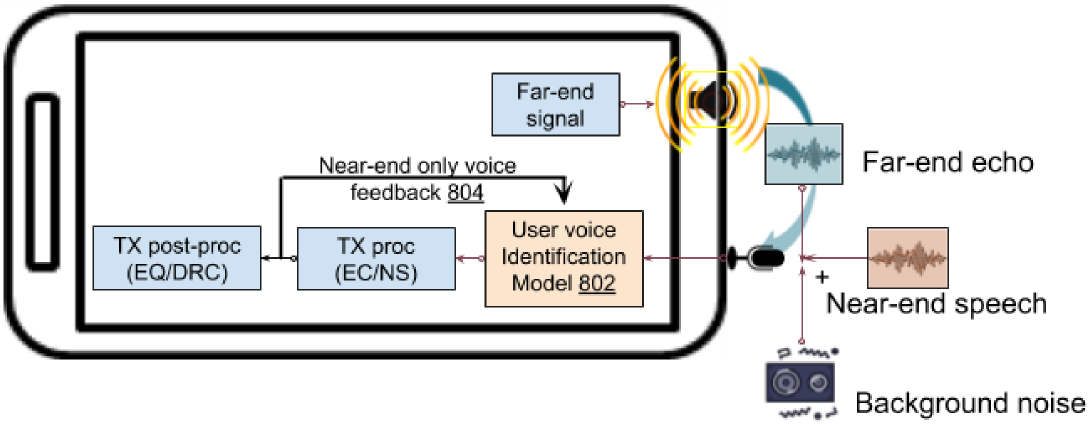

Fig. 8: Training the voiceprint model in an online manner

Fig. 8 illustrates the training of the voiceprint model in an online manner. Features are provided (e.g., via an API) that enable the user to provide speech input by reading a paragraph. The user provides the speech input at a time before the start of a call. The speech input is used to build a baseline user-specific voiceprint model (802). During a call, the voiceprint model updates using voice data (804) that is obtained after echo cancellation and noise suppression. Using voice data that is free of far-end echoes and background noise enables training of the model with nearly pure user speech.

In this manner, a user's voiceprint, which is a unique set of voice features detected using machine learning models, is used to distinguish the user's speech from surrounding disturbances, and to exclusively enhance the user's speech.

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 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 techniques that use the voiceprint, which is a set of features unique to a user’s voice, to home into the user’s speech and cut out surrounding disturbances. The techniques are implemented with user permission to generate and use the voiceprint.