November 21, 2017

Acoustic echo cancellation using sub-band encoded training sequences

Milton Ribeiro Filho

Acoustic echo cancellation using sub-band encoded training sequences

ABSTRACT

An acoustic echo canceler (AEC) enhances speech quality by removing echoes. An AEC operates by estimating the acoustic path, e.g., transfer function, between a speaker and a microphone. A training sequence of symbols, known a priori, is transmitted to probe the acoustic environment between the speaker and microphone and used to determine the transfer function. The transfer function is used to cancel echoes from the signal produced by the microphone. The performance of an AEC is affected by various factors such as a dynamic acoustic path, multiple acoustic paths, presence of background noise, etc. In certain conditions, the time required for robust estimation of the transfer function can be such that the transfer function changes during the estimation. Repeated transmittal of the training sequence can improve transfer function estimates; however, the training sequence itself occupies bandwidth that can otherwise carry speech payload.

Techniques of this disclosure achieve robust AEC performance by transmitting pseudo-random training sequences at spectral regions unoccupied by speech, in a manner imperceptible to the human ear. The transfer function can be estimated and updated nearly continuously and imperceptibly.

KEYWORDS

- Acoustic echo cancellation
- Sub-band coding
- Transfer function estimation
- Training sequence

BACKGROUND

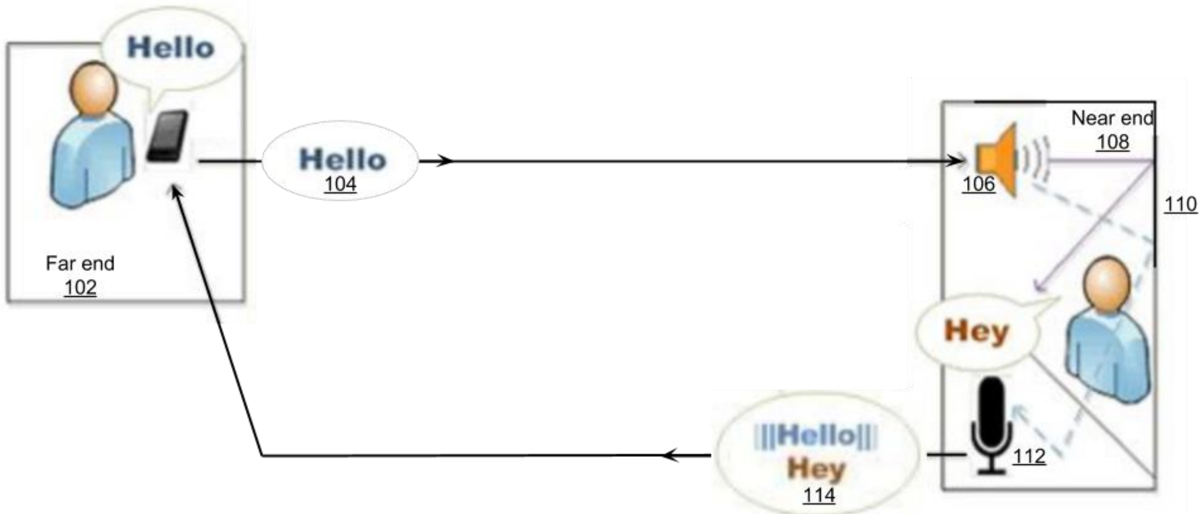

Fig. 1: Formation of acoustic echoes in a telephone call

Fig. 1 illustrates the formation of acoustic echoes in voice communication, e.g., a telephone call. At the far end (102), a participant in the call says “Hello.” The speech is transmitted (104) to the near end (108), where it is output by a speaker (106). Aside from reaching human ears at the near end, the output from the speaker also reflects off walls and objects (110) at the near end. Microphone (112) detects the near-end participant’s speech signal (“Hey”) along with the reflections. The detected signal (114) is sent to the far end.

Acoustic echo cancelers estimate the acoustic path, e.g., the multipath profile or transfer function, between the speaker and the microphone. Once the acoustic path is known, an echo is canceled by synthesizing a negative echo at the delay of the echo. To estimate the acoustic path, training sequences are inserted into the audio stream. Training sequences occupy bandwidth that can otherwise usefully carry speech payload, and hence cannot be inserted indiscriminately.

The acoustic path is generally non-stationary, e.g., it varies with time as the relative positions of reflective objects in a room change. Therefore, to achieve echo cancellation, repeat transmittal of the training sequence are necessary. The non-stationarity of the acoustic path can also cause a lack of convergence in the transfer function estimate, e.g., when the acoustic path changes before the estimation is completed. The transfer function estimate also fails to converge if there is high correlation between sounds from multiple sources, e.g., when sounds are output by a stereo speaker systems. Even in the case of relatively stationary acoustic paths or mono-speaker systems, transfer function estimates are rendered less-than-accurate by background noise, multiple acoustic paths, etc.

Sub-band coding (SBC) is a signal processing technique frequently used in lossy audio compression algorithms such as MP3. SBC exploits auditory masking in the human auditory system. Human ears are normally sensitive to a wide range of frequencies; however, when a sufficiently loud signal is present at one frequency, the ear does not detect weaker signals at nearby frequencies since the louder signal masks the softer ones. The louder signal is called the masker, and the point at which masking occurs is known as the masking threshold.

DESCRIPTION

This disclosure applies the principle of loud signals masking soft signals at nearby frequencies to the problem of acoustic echo cancellation as follows. Per techniques described herein, training sequences are inserted at spectral regions not currently occupied by speech, at a level below masking threshold. The training sequences are imperceptible to the call participants and enable estimation of the acoustic path nearly continuously.

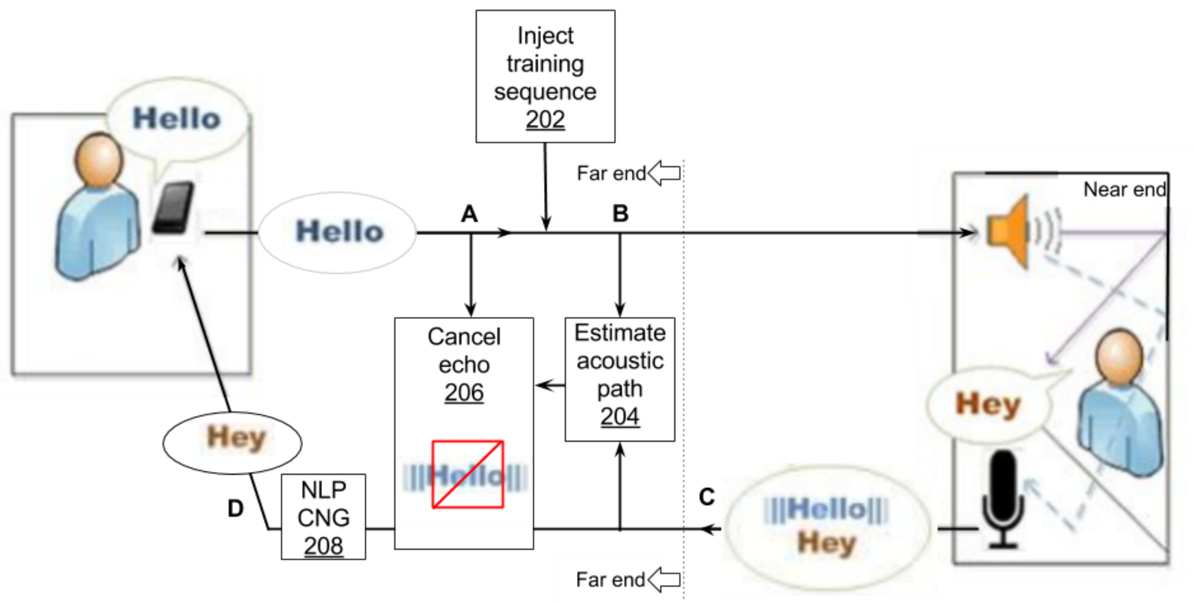

Fig. 2: Acoustic echo canceler with below-masking-threshold training sequence

Fig. 2 illustrates an acoustic echo canceler per techniques of this disclosure. A training sequence is injected (202) into the signal that carries the far-end speech. This training sequence is below the masking threshold of the speech originating at either end. The training sequence is a wide-band, pseudo-random sequence with a flat frequency spectrum. The training sequence is injected into regions of the spectrum where speech-signal energy is low or absent. The instantaneous location in frequency of the training sequence depends on the current spectral occupancy pattern of the speech signal, and varies with time.

Both far-end speech and training sequence are received at the near end and output by a speaker. Echoes of these signals are detected by the near-end microphone. When the composite signal that includes the near-end speech and far-end echoes reaches the far end, the acoustic path (transfer function) between the near-end speaker and the near-end microphone is estimated (204) by comparing the received and transmitted training sequences.

The acoustic path estimator (204) is an adaptive filter whose taps are adapted by algorithms such as least-mean squares (LMS), recursive least squares (RLS), etc. The acoustic path estimator filters out the received training sequence prior to using it in an adaptive training algorithm. This is possible since the instantaneous location in frequency of the transmitted training sequence is known. The acoustic path estimator outputs an acoustic path estimate, which is a digital filter that models the acoustic environment at the near end.

The acoustic path estimate is used to synthesize at the far end a negative copy of the echo which is used to cancel the echo generated at the near end (206). Synthesis of the echo is achieved by passing a delayed copy of the far-end speech signal through the acoustic path estimate. The bulk of the echo is removed. However, small traces may remain. These are removed and/or masked (208) by the non-linear processor (NLP) and the comfort-noise generator (CNG).

To explain the relative positions (in frequency) and amplitudes of the training sequence vis-a-vis the speech signals, Fig. 3 illustrates spectra at various points along the communication path.

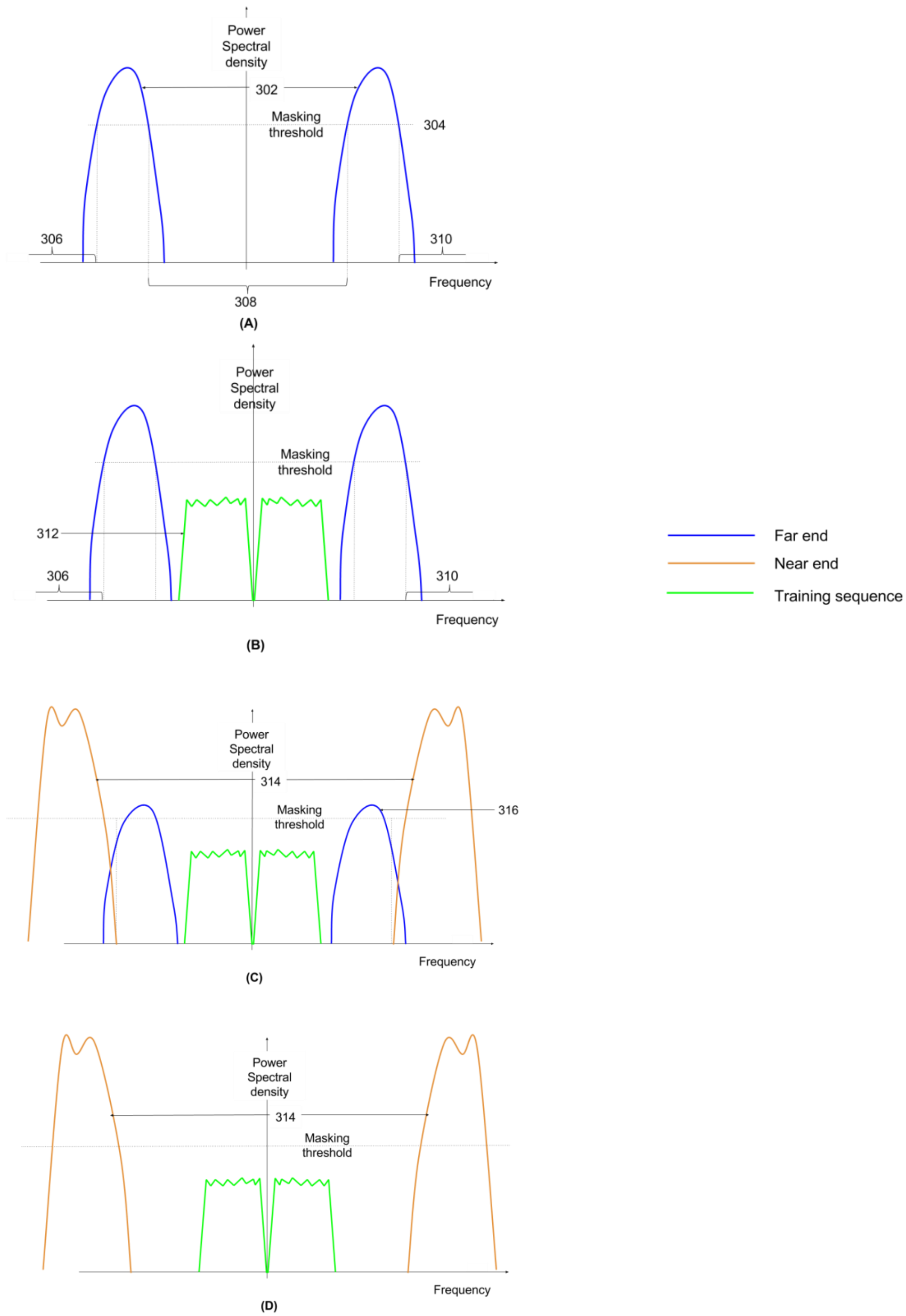

Fig. 3: Spectra of the near end, far end and training sequences at various points in the communication path

Fig. 3(A) shows spectrum as the speech signal emerges out of the far end (point A of Fig. 2). At this point, the spectrum includes only the far-end speech signal (302). Spectral energy below masking threshold (304) is not perceptible to the human ear. Regions of the spectrum 306-310 have far-end spectral energy either below threshold or absent.

Fig. 3(B) shows spectrum after the training sequence is injected (point B of Fig. 2). The training sequence (312) is inserted in unoccupied spectral regions. The training sequence is of low enough power that it stays below masking threshold. The spectral density of the training sequence is almost flat.

Fig. 3(C) shows spectrum at the point of emergence from the near-end microphone (point C of Fig. 2). At this point, the spectrum comprises near-end speech (314), echoes of the far-end speech (316), and the training sequence. The echoes of the far-end speech are above the masking threshold, rendering the echoes audible (in the absence of AEC) at the far-end.

Fig. 3(D) shows the spectrum after echo cancellation (point D of Fig. 2). The far-end echoes are removed, and only the near-end signal (314) and the training sequence remain. The training sequence, below masking threshold, is imperceptible.

In case of stereo sound at an end, more than one speaker, and/or more than one microphone is present at that end. In this case, the transfer function for every speaker-microphone pair is estimated. A distinct training sequence is transmitted for each speaker-microphone pair. The multiple training sequences used for stereo sound are orthogonal to each other, such that they can be separated for the purposes of estimating the multiple transfer functions. The orthogonality of the training sequences can be achieved in different ways, e.g., by using zero- or low-correlation pseudo-random sequences, by using sequences that have spectral footprints that cover distinct and unoccupied portions of the spectrum, etc. The techniques of this disclosure can be used standalone or be combined with other acoustic echo cancellation techniques.

CONCLUSION

Techniques of this disclosure achieve acoustic echo cancellation in situations where the acoustic path is not static, when there are multiple acoustic paths, when background noise is present, where multiple speakers are used, etc. A training sequence is injected in spectral regions unoccupied by speech signals. The training sequence is at a power below masking threshold and is imperceptible to human ears. Using the training sequence, the acoustic echo canceler continuously estimates the acoustic path in a fast-converging manner, enabling echo cancellation.