In recent years, people that are interested in full-duplexed, multichannel sounds for telepresence services have realized that the acoustic echo cancellation (AEC) problem is mathematically ill-conditioned. To regularize the problem, it is known that signals in multiple channels must be decorrelated. In this paper, we propose to utilize decorrelated background sounds for solving the AEC problem. Methods are presented to generate arbitrarily many orthogonal but perceptually similar copies from a mono sound, and their effectiveness in 2-channel and 5-channel AEC is evaluated in simulations.

INTRODUCTION

With the availability of increased communication bandwidths in recent years, people have become interested in full-duplexed, multichannel sounds for telepresence services because of the potentials for providing much better hearing experiences. Nevertheless, in any full-duplex connection of audio network, the problem of acoustic echoes arises due to the coupling between loudspeaker(s) and microphone(s) placed in the same room. Moreover, it is known [1][2] that the cancellation of multichannel acoustic echoes is a mathematically ill-conditioned problem. What happens is that, due to the high correlation between signals in multiple channels, an adaptive echo canceler tends to converge to a degenerate solution and fails to find the true coupling paths.

Although several types of algorithms for decorrelating the channels [1]-[6] have been proposed to regularize the problem in the context of speech teleconferencing, these algorithms are all developed based on the criterion that any sound other than the speech signals should not be heard. However, it is not necessarily so in applications such as video games and performing arts where background sound effects and background music are common and desired practices. In this paper, we propose to utilize background sounds for multichannel echo canceling. Methods are developed to generate arbitrarily many orthogonal and perceptually similar sounds from a mono source, and the sounds are fed into a multichannel echo canceler for the canceler to better identify the echo paths.

The rest of the paper is organized as follows. Section 1 gives a literature review on the ill-conditioning of the multichannel AEC problem. Also, it is briefly pointed out why the problem can be regularized by inter-channel decorrelation. Section 2 describes the proposed methods to generate perceptually similar orthogonal sounds. Experiments evaluating the effectiveness of the orthogonal sounds for AEC are documented in Section 3, and future research directions are discussed in Section 4.

1. LITERATURE REVIEW

1.1. The non-uniqueness problem of multichannel acoustic echo canceling

In a multichannel teleconferencing setup, the nonuniqueness problem arises from the fact that the signals in the multiple channels are linearly filtered from the same source.

Figure 1: Schematic diagram of multichannel echo cancellation. In this case, only two channels are shown.

Without loss of generality, here we review the two-channel (i.e., “stereophonic”) case. Assuming the configuration as shown in Fig. 1, we have

\[ x_i(n) = s(n) * g_i(n), \quad i = 1,2 \]

where \( s(n) \) is the source signal, \( x_1(n) \) and \( x_2(n) \) are the signals in the two channels, the \(*\) sign denotes the convolution operator, and \( g_1(n) \) and \( g_2(n) \) are the coupling impulse responses from the source to the two microphones in the recording room. It is straightforward to see that the two
channels are linearly dependent as described by the following relation,

\[ x_1(n) * g_2(n) = x_2(n) * g_1(n) \] (1)

Therefore, any identified impulse response \( \{h_1(n), h_2(n)\} \) that can be written as the true response \( \{h_1(n), h_2(n)\} \) plus a degenerate term \( \{\beta(n) * g_2(n), -\beta(n) * g_1(n)\} \) is a valid echo canceling solution, where \( \beta(n) \) is an arbitrary time-sequence. This can be easily verified by the following steps,

\[
\hat{y} = x_1 * \hat{h_1} + x_2 * \hat{h_2} \\
= x_1 * (h_1 + \beta * g_2) + x_2 * (h_2 - \beta * g_1) \\
= (x_1 * h_1 + x_2 * h_2) + \beta * (x_1 * g_2 - x_2 * g_1) \\
= y
\]

We have just derived that \( e(n) = y(n) - \hat{y}(n) = 0 \), i.e., the echo is perfectly canceled, although there is a mismatch between the identified response and the true response. One may falsely conclude that multichannel echo canceling is hence more easily achieved than single channel echo canceling because any of the infinitely many degenerate solutions \( \{\hat{h}_1, \hat{h}_2\} \) works. However, this is a wrong conclusion because at a typical teleconference, many people in a room can take turns talking and cause abrupt changes in \( \{g_1(n), g_2(n)\} \). Every time this occurs, it takes a learning time for the echo canceler to converge to a new degenerate solution. Consequently, the echo canceler doesn’t practically work unless the mean duration each turn a person talks is much greater than the learning time of the echo canceler, which is typically on the order of 1-10 seconds.

1.2. Ill-conditioning

In practice, when the length of the modeling filters (\( L \)) of the echo canceler is shorter than the length of the room response (\( N \)), there is a great mismatch between the identified response and the true response when the signals in various channels are highly correlated. Let the impulse response from the \( i^{th} \) loudspeaker to one of the \( J \) microphones be expressed as follows,

\[
h_{i,N} = \begin{bmatrix} h_{i,L} \\ \hat{h}_{i,t} \end{bmatrix}
\]

where \( h_{i,N} \) is the actual room response, \( h_{i,L} \) is the first \( L \) samples of it, and \( \hat{h}_{i,t} \) is the tail of the response. The mismatch between the identified response and the true room response can be quantified by the misalignment, defined as

\[
e = ||h - \hat{h}||/||h||
\] (2)

where \( h = [h_{1,L}^T, ..., h_{J,L}^T]^T \) is the J-channel truncated true response, and \( \hat{h} = [\hat{h}_{1,L}^T, ..., \hat{h}_{J,L}^T]^T \) is the identified response.

It is shown in [1] that the minimum mean square error of the misalignment is given by

\[
e_{\text{min}}^2 = \frac{h_i^T Q_i h_i}{h_i^T h_i}
\] (3)

where

\[
Q_i = R_i^T R_i^{-2} R_i
\]

Here, \( R \) is covariance matrix of the size \( JL \times JL \) and \( R_i \), which has the size \( JL \times J(N - L) \), is the tail of covariance. It is now apparent that if \( R_i^{-1} \) is ill-conditioned, a large misalignment will occur as a consequence. To regularize the covariance matrix and reduce the misalignment, it is shown in [1] that the signals \( x_j \) in different channels need to be decorrelated. It is proved in [1] that the eigenvalues of the covariance matrix \( R \) are bounded below by a factor \([1 - |\gamma_{ij}(f)|]^2\), where

\[
\gamma_{ij}(f) = \frac{S_{x_i,x_j}(f)}{\sqrt{S_{x_i,x_i}(f)S_{x_j,x_j}(f)}}
\] (4)

is the coherence spectrum, defined as the normalized Fourier transform of the cross-correlation function between \( x_i \) and \( x_j \). Qualitatively, when \( x_i \) and \( x_j \) are decorrelated, \( \gamma_{ij}(f) \) is low at all frequencies, \( R_i^{-1} \) is well-behaved, and the misalignment is low. I.e., inter-channel decorrelation helps the adaptive echo canceler to identify the true response. However, any of such decorrelation algorithms has to satisfy the constraint that it introduces no audible artifacts or distortions. In the next section, we will review some of the decorrelation algorithms and discuss their advantages and shortcomings.

2. METHODS

2.1. Choices of interchannel decorrelation methods

In designing a method to decorrelate background sounds that can be used in arbitrarily many channel echo cancelation, the following issues need to be considered. First, a method of decorrelation, as defined in [7], should introduce no perceptible distortion to the original signals. Second, the method should maximally regularize the J-channel covariance matrix \( R \). The ideal case of this would be to orthogonalize signals in different channels:

\[
\lim_{M \to \infty} \frac{1}{M} \sum_{n=0}^{M} x_i(n)x_j(n - m) = 0, \quad i \neq j, \quad \forall m
\] (5)

or equivalently,

\[
S_{x_i,x_j}(f) = 0, \quad i \neq j.
\] (6)
While many different decorrelation methods have been proposed and evaluated [1] - [6] for speech communication in stereophonic teleconferences, not all of them can be readily applied to generate arbitrarily many orthogonal signals. These methods are briefly summarized in Table 1.

Within the scope of this paper, the main concern is to decorrelate arbitrarily many channels. Therefore, the three applicable types of methods mentioned in Table 1 are used to design the proposed methods that will be described and evaluated in the following sections of this paper. These proposed methods can be further categorized as either “partial orthogonalization” or “total orthogonalization” based on whether or not the coherence spectrum $S_{\pi,\pi}(f)$ is suppressed all the way to 0.

2.2. Partial orthogonalization by watermarks

We implemented an interchannel decorrelation scheme that embeds orthogonal digital watermarks [8] in signals in arbitrarily many channels. The scheme is similar to [3] except that a simpler masking model is adopted. The model uses a two-slope approximation to the spreading function [9] that is illustrated in Fig. 2, where the spikes represent sinusoidal tones, and the shaded area indicates how the watermark energy can be allocated. The empirical formula for the two slopes are given as follows,

$$S_{\text{left}} = 27 \text{ dB/Bark};$$

$$S_{\text{right}} = \begin{cases} -27, & \text{if } \Gamma < \Gamma_{\text{th}}; \\ -27 + 0.37 \cdot (\Gamma - \Gamma_{\text{th}}), & \text{otherwise}, \end{cases}$$

where $S_{\text{left}}$ and $S_{\text{right}}$ are the two slopes in units of dB/Bark, $\Gamma$ is the magnitude in dB of the masker tone, and $\Gamma_{\text{th}}$ is an empirical threshold.

In addition to the two slopes, the parameter $\Delta$ as shown in Fig. 2 controls how much energy of watermarks is to be injected. Informal listening suggests that a $\Delta$ of 33dB or more preserves the perceptual quality of the host signals pretty well, while a $\Delta$ of less than 25dB makes the watermarks audible and disturbing.

2.3. Total orthogonalization: time-varying all-pass filtering

For background sounds such as music, we propose to use time-varying all-pass filters [6] to generate perceptually similar orthogonal sounds. The generation of such sounds involves the generation of a Hilbert transform pair and orthogonal time-varying linear transformations of the pair,

$$x_j(n) = \cos(\omega_j n T)x_0(n) + \sin(\omega_j n T)\tilde{x}_0(n)$$

where $x_0(n)$ is the mono signal background music whose perceptually similar copies are to be panned to arbitrarily many channels, $\{x_0(n), \tilde{x}_0(n)\}$ forms a Hilbert transform pair (i.e., analytic signal), and $\{\cos(\omega_j n T), \sin(\omega_j n T)\}$ represents slow phase modulators. The perceptual similarity between different $x_j$ is based on the fact that human listening is insensitive to small phase shifts. Also, the orthogonality holds as long as $\omega_j$ are all different and nicely spaced from one another. This claim can be explained: since $\tilde{x}_0(n)$ is 90-degrees phase-shifted from $x_0(n)$ at all frequencies, $\{x_0(n), \tilde{x}_0(n)\}$ can be thought of as two orthogonal basis vectors in the space they span. Conceptually, equation (7) gives signals
that are rotating around the origin in this space at different angular velocities \( \omega_j \). The orthogonality condition \( S_{n,x_j}(f) = 0 \) holds in long term if \( \omega_i \neq \omega_j, \forall i \neq j \).

Note that the modulating terms \( \cos(\omega_j n T) \) and \( \sin(\omega_j n T) \) cause a frequency shift of \( \omega_j \) in all frequency bands. In order to avoid introducing audible distortion in \( x_j(n) \) as compared against \( x_0(n) \), we choose \( \omega_j < 2 \cdot 2\pi \text{ rad/s} \). Although this prevents audible artifacts as far as each individual channel is concerned, as reported in [2], the time-varying phase differences between signals in multiple channels cause a joint artifacts that can be interpreted as time-varying directions of arrival of the sounds. Although our informal listening also confirms the existence of such joint-channel artifacts, we remain optimistic about the application of the total orthogonalization algorithm on background sound effects, but not on foreground speech signals.

We have experimented with this total orthogonalization algorithm on both wideband noise-like background sound effects and tonal background music. Coherence spectra of two selected cases are shown in Fig. 3 and Fig. 4, with angular velocities \( \omega_j \) of \( 0, \frac{\pi}{3}, 1, 1.5, \) and \( 2 \text{ Hz} \). The original signal \( x_0(n) \) is a white-noise for the case shown in Fig. 3, and a clip from the opening part of Scriabin’s piano sonata No. 1 for the case shown in Fig. 4. The spectra are calculated using a length 128 FFT (at the sampling rate of 16kHz) of the cross correlation functions that are estimated by 15 seconds of averaging. Note that both figures illustrate that \( \gamma_{ij}(f) \) are bounded above by 1 as expected, whenever \( i \neq j \). However, the estimated coherence spectra are not exactly zero due to a finite time of averaging.

Figure 3: Interchannel coherence spectra of total-orthogonalized noise-like background sounds. Only 5 of the spectra are shown to save space.

Figure 4: Interchannel coherence spectra of total-orthogonalized tonal background music.

3. EXPERIMENTS

3.1. Stereophonic AEC simulation: decorrelation by watermarking

Fig. 5 shows the result of a stereophonic AEC simulation. Misalignment versus time is shown at various levels of watermarking. The original signals are a male voice speaking the sentence “The fifth jar contains big, juicy peaches,” and are identical in the two channels before the injection of masked noise using the partial orthogonalization method. The \( \Delta \) that controls the energy of watermarks is always set to be 30dB or more so that the watermarks can hardly be heard, if not inaudible. The virtual room of the simulation has a size of 4m (width) by 5m (length) by 3m (height), and both loudspeakers are placed within 3 meters from a microphone at the position \((2.0, 2.0, 1.5)\). The coupling impulse responses are calculated using the image method [10], and then truncated to the length of 320 at the sampling rate of 16kHz. The adaptive algorithm for AEC is generically chosen to be \( \alpha \)-LMS [11] with an \( \alpha \) of \( \frac{5}{8} \).

In Fig. 5, note that without decorrelation, the misalignment converges rapidly but stays at about 50%. The injection of watermarks successfully reduces the misalignment beyond that 50% limit and helps the echo canceler to better identify the impulse response. The result of decorrelation by halfwave rectification is also shown for comparison purposes.

3.2. 5-channel AEC simulations: decorrelation by total orthogonalization

Fig. 6 illustrates the setup of a 5-channel AEC simulation in a virtual room of the size 10m \( \times \) 10m \( \times \) 3m. The
sound reflection ratios of the walls, the floor, and the ceiling are set arbitrarily between 0.5 and 0.8. For simplicity, the loudspeakers and the microphones are assumed omnidirectional.

Two different classes of orthogonalized background sounds are tested using the following original sounds respectively,

- "ocean wave": synthesized by slowly modulating the amplitude of a noise signal recorded in a typical office environment with 60Hz and harmonics rejected.
- "piano music": the Piano Sonata No. 1, 1st movement, composed by A. Scriabin.

Fig. 7 compares the misalignment versus time of these two cases to the case of no decorrelation. The filter length $L$ is 1600 and the coupling response length $N$ is 8192, at the sampling rate of 16kHz. It is obvious that both classes of orthogonalized sounds successfully achieve a significant percentage of reduction of the misalignment, as compared against the case of no decorrelation.

Also, it is shown that more than 30dB of echo cancellation can be obtained toward the end of a 1-minute AEC simulation. Fig. 8 shows the actual waveforms during the last 15 seconds of the simulation. The original “piano music” $x_0(n)$, the signal $y_2(n)$ recorded back by the front center channel before AEC, and the cancellation error $e_2(n)$ after AEC are shown from top to bottom respectively. In this particular simulation, a cancellation ratio of 32.2dB is achieved.
4. FUTURE DIRECTIONS

In this paper, we propose two different types of methods to generate arbitrarily many orthogonalized copies of perceptually similar background sounds. We also have experimented with the methods separately in the case of 2-channel and 5-channel AEC. However, we have not tried to jointly use these two methods to see if the performance can further be improved.

Also, through the AEC experiments we have conducted, we found that there are tunable parameters in both the orthogonalization stage and the adaptive learning stage. In the orthogonalization stage, which is the main concern of this paper, the tunable parameters include $\Delta$, which controls the energy of watermarks; and $\omega_j$, the frequency shifts. While the parameters reported in this paper are selected in an ad hoc way, we are interested in formulating the selection as an optimization problem and would like to pursue this direction in the future.

5. CONCLUSIONS

We have proposed to use background sounds for arbitrarily many channel acoustic echo canceling. Techniques of generating orthogonal but perceptually similar copies of the same sound were developed, analyzed, and evaluated. The 2-channel and 5-channel simulations we conducted support that perceptually similar orthogonal sounds successfully regularize the multichannel AEC problem and achieve cancellation of acoustic echoes.

REFERENCES


