Frequency-Domain Double-Talk Detection Based on the Gaussian Mixture Model
Kyu-Ho Lee, Joon-Hyuk Chang, Member, IEEE, Nam Soo Kim, Member, IEEE, Sangki Kang, and Yongserk Kim

Abstract—In this letter, we propose a novel frequency-domain approach to double-talk detection (DTD) based on the Gaussian mixture model (GMM). In contrast to a previous approach based on a simple and heuristic decision rule utilizing time-domain cross-covariances, GMM is applied to a set of feature vectors extracted from the frequency-domain cross-correlation coefficients. Performance of the proposed approach is evaluated through objective tests under various environments, and better results are obtained as compared to the time-domain method.

Index Terms—Cross-correlation coefficient, double-talk detection, Gaussian mixture model, likelihood, voice activity detector.

I. INTRODUCTION

RECENT advance in communication technologies has encouraged the usage of small hands-free devices, which provides more convenience for mobile communication users. However, in compact hands-free terminals, the acoustic echo that arises due to coupling between a loudspeaker and microphone is unavoidable. Particularly when the echo path changes rapidly, it becomes more difficult to estimate the acoustic echo path response. A traditional solution for this problem is an acoustic echo canceller (AEC) system [1], [2]. Most of the traditional AEC algorithms are built on an adaptive finite impulse response (FIR) filter for identifying the echo path response, and subtract the echo estimate from the microphone input signal. In those AEC systems, double-talk detection (DTD) is a core component since the near-end signal acts as a source of disturbance in updating the filter coefficient. Recently, DTD algorithms based on the cross-correlation, signal-level, and coherence were proposed, showing enhanced detection performances [3]–[7]. Among these, the cross-correlation-based method is known to provide impressive performance with low computational complexity. However, it requires a voice activity detection (VAD) module for far-end speech, and performance degrades in the presence of background noise or when rapid echo path change occurs because the residual error of the echo canceller significantly affects the DTD performance due to the jointly coupled structure.

In this paper, we first consider frequency-domain acoustic echo suppression (AES) and then analyze the components that contribute to the computation of cross-correlation. After in-depth analysis of the contributing components, a feature vector that is useful for discrimination between double-talk and single-talk is then selected by approximating each class conditional distribution with a Gaussian mixture model (GMM). The proposed approach can be efficiently implemented in the frequency domain, and experimental results show that it yields better results as compared to the conventional cross-correlation-based method.

II. CORRELATION BASED DTD ALGORITHM

In this section, we briefly review the cross-correlation based DTD algorithm as given in [7]. In the time domain, acoustic echo cancellation is performed by using an FIR adaptive filter. Let \( x(k), \hat{w}(k), \hat{d}(k), \) and \( z(k) \) denote the far-end signal, the echo path response, the estimated echo, and the echo path output, respectively. Then, we have

\[
\hat{d}(k) = \hat{w}^T(k)\hat{x}(k) \tag{1}
\]

\[
\hat{c}(k) = z(k) - \hat{d}(k) \tag{2}
\]

\[
\hat{w}(k+1) = \hat{w}(k) + \beta \frac{\hat{c}(k)}{P_{\hat{c}}(k)+\gamma} \hat{x}(k)e(k) \tag{3}
\]

where \( \hat{w}(k) = [\hat{w}_0(k), \hat{w}_1(k), \ldots, \hat{w}_{N-1}(k)]^T \) is the \((N \times 1)\) filter coefficient vector, \( \hat{x}(k) = [x(k), x(k-1), \ldots, x(k-N+1)]^T \) is the \((N \times 1)\) input vector, \( T \) denotes matrix transposition, \( \beta \) is a factor controlling the convergence, \( P_{\hat{c}}(k) \) is the power of the input signal and \( \gamma \) is a stabilization factor. Under this AEC framework, two types of cross-correlations are considered at time \( k \). First, the cross-correlation coefficient between the microphone output and the estimated echo is defined by

\[
\phi_{ze}(k) = \frac{p_{ze}(k)}{\sqrt{p_e(k)p_z(k)}} \tag{4}
\]

where \( p_e(k) \) is the power of the estimated acoustic echo signal, \( p_z(k) \) is the power of the microphone signal, and \( p_{ze}(k) \) is the cross-power between the two. Second, the cross-correlation coefficient between the microphone input and the residual error of echo canceller is given by

\[
\phi_{zd}(k) = \frac{p_{zd}(k)}{\sqrt{p_z(k)p_d(k)}} \tag{5}
\]

where \( p_d(k) \) is the power of the error signal, and \( p_{zd}(k) \) is the cross-power between the microphone and error signals. In [7], it is discovered that the value of \( \phi_{zd}(k) \) is close to 1 for single-talk periods and approaches 0 for double-talk periods, which explains why \( \phi_{zd}(k) \) is a good indicator for single-talk.

Authorized licensed use limited to: Seoul National University. Downloaded on April 14,2010 at 09:12:38 UTC from IEEE Xplore. Restrictions apply.
other hand, \( \phi_{2D}(k) \) is a useful indicator for double-talk since its value is close to 1 during double-talk periods. In practice, these values are used in a complementary way for detecting the double-talk periods [7].

III. DTD BASED ON GMM

In this work, we focus on frequency-domain rather than time-domain approaches to AES, due to their good performance with low computational complexity [8]. We also apply the GMM to classify multidimensional features extracted in the frequency domain. Let \( Z(t, i), \hat{D}(t, i) \), and \( E(t, i) \) denote the \( i \)th discrete Fourier transform (DFT) coefficient of the microphone signal, the estimated acoustic echo, and the residual echo signal, respectively, at the \( t \)th frame. Similar to the way we first compute the cross-correlation in the time domain, we define two cross-correlation functions, \( \Phi_{ZD}(t, i) \), and \( \Phi_{ZE}(t, i) \) as follows:

\[
\Phi_{ZD}(t, i) = \frac{P_{ZD}(t, i)}{\sqrt{P_Z(t, i)P_D(t, i)}}
\]
\[
\Phi_{ZE}(t, i) = \frac{P_{ZE}(t, i)}{\sqrt{P_Z(t, i)P_E(t, i)}}
\]

where \( P_{ZD}(t, i) \) and \( P_{ZE}(t, i) \) are the cross power spectra defined as \( z(k), \delta(k) \), and \( e(k) \) for the \( i \)th frequency bin computed at frame \( t \). Similarly, it is easy to understand that \( P_Z(t, i) \), \( P_D(t, i) \), and \( P_E(t, i) \) denote the powers of \( z(k), \delta(k) \), and \( e(k) \) for frequency bin \( i \) at the \( t \)th frame. Practically, all the auto and cross power spectra needed in (6) are estimated as follows:

\[
P_{ZD}(t, i) = \alpha_p P_{ZD}(t-1, i) + (1 - \alpha_p) |Z(t, i)\hat{D}^*(t, i)|^2
\]
\[
P_D(t, i) = \alpha_p P_D(t-1, i) + (1 - \alpha_p) |\hat{D}(t, i)|^2
\]
\[
P_Z(t, i) = \alpha_p P_Z(t-1, i) + (1 - \alpha_p) |Z(t, i)|^2
\]
\[
P_{ZE}(t, i) = \alpha_p P_{ZE}(t-1, i) + (1 - \alpha_p) |Z(t, i)E^*(t, i)|^2
\]
\[
P_E(t, i) = \alpha_p P_E(t-1, i) + (1 - \alpha_p) |E(t, i)|^2
\]

where \( \alpha_p \) is a smoothing constant and \( * \) denotes the complex conjugate. In order to verify whether \( \Phi_{ZD}(t, i) \) and \( \Phi_{ZE}(t, i) \) are applied as useful features for DTD, we first obtained histograms (distributions) of these cross power spectra over a set of training data consisting of both single-talk and double-talk periods. After an extensive analysis of the spectrum obtained from speech data as shown in Fig. 1 as an illustrative example, we selected only seven features including \( \Phi_{ZD}(t, i) \) for frequency bin \( i = 1, 2, 3 \) and \( \Phi_{ZE}(t, i) \) for frequency bin \( i = 1, 3, 5, 6 \), because we should consider both the statistical discrimination capability and the additional computation cost as in [9]. Actually, it is observed that the selected features clearly accounts for the statistical discrimination between single-talk and double-talk. Accordingly, we have come to the conclusion that the selected features distribution for each state, i.e., single-talk or double-talk, can be well characterized by a GMM due to the multimodal characteristics. The Gaussian mixture density is defined by a weighted sum of \( M \) component Gaussian densities for the feature vector \( \Phi \) as follows:

\[
p(\Phi|\lambda) = \sum_{j=1}^{M} \alpha_j p_j(\Phi)
\]

where

\[
\sum_{j=1}^{M} \alpha_j = 1,
\]

\[
p_j(\Phi) = \frac{1}{(2\pi)^{d/2} \left| \Sigma_j \right|^{1/2}} \exp \left\{ -\frac{1}{2} (\Phi - \mathbf{m}_j)^T \Sigma_j^{-1} (\Phi - \mathbf{m}_j) \right\}
\]

Here, \( \alpha_j \) is the weight for the \( j \)th mixture of the GMM and \( \mathbf{m}_j \) and \( \Sigma_j \) represent the corresponding mean vector and covariance matrix, respectively. These parameters are collectively denoted by

\[
\lambda = \{ \alpha_j, \mathbf{m}_j, \Sigma_j \}, \quad j = 1, \cdots, M.
\]

For single-talk and double-talk classification, the data distribution for each class is approximated by a GMM, i.e., the single-talk model \( (\lambda_s) \) and the double-talk model \( (\lambda_d) \) for which the maximum likelihood (ML) estimation is achieved to estimate the relevant parameters of the GMMs. For convenience of implementation, we apply diagonal covariance matrices in our algorithm. Given the trained single-talk and double-talk models, the input frame is classified into either single-talk or double-talk, based on the likelihood ratio (LR) test:

\[
\log \Lambda(t) = \log \frac{p(\Phi(t)|\lambda_d)}{p(\Phi(t)|\lambda_s)} \geq \eta
\]

where \( \eta \) is the experimentally selected detection threshold. It should be noted that a smoothed LR is desired to prevent the frequent transition from one state to another by considering inter-frame correlation of speech. Let, \( \psi(t) \) denote the smoothed LR computed at frame \( t \). Then

\[
\psi(t) = \exp \left[ \beta_p \log \psi(t - 1) + (1 - \beta_p) \log \Lambda(t) \right]
\]

where \( \beta_p \) is an experimentally chosen smoothing factor. The proposed DTD algorithm can be incorporated into the conventional AES system as shown in Fig. 2. From Fig. 2, it can be seen that the DTD part controls the update of echo path response \( \hat{W}(t, i) \) for freezing in the case of double-talk. By following Faller’s method proposed in [10], AES is achieved through

\[
E(t, i) = G(t, i) Z(t, i)
\]

where \( G(t, i) \) is a parametric Wiener filter (PWF) gain given by

\[
G(t, i) = \left\{ \frac{\max \left[ \left| Z(t, i) \right|^2 - \nu \left| \hat{D}(t, i) \right|^2, 0 \right] }{ \left| Z(t, i) \right|^2 } \right\}^{1/2}
\]

in which \( \left| \hat{D}(t, i) \right| \) is substantially obtained by the magnitude of the least squares estimator [10] as follows:

\[
\left| \hat{D}(t, i) \right| = \hat{W}(t, i) |X(t, i)|
\]
where
\[
\hat{W}(t,i) = \frac{E[X^*(t,i)Z(t,i)]}{E[X^*Z(t,i)]^2}.
\]

Also, \(\xi\) and \(\nu\) are control parameters in echo suppression that makes a compromise between the speech distortion and residual echo.

**IV. EXPERIMENTAL AND RESULTS**

The performance of the proposed DTD algorithm was measured in terms of the DTD error probability \(P_E\), which is the sum of the false-alarm and miss probabilities. We compared \(P_E\) to that of the time-domain cross-correlation based method proposed in [7]. For this, utterances from two male and two female speakers were used to construct 768-s of speech data for training. We artificially created 20 data files, where each file was obtained by mixing the far-end signal with the near-end signal. As we mentioned in the previous section, each frame of the windowed signal was converted into a 128-point DFT after zero padding. The far-end speech signal was passed through a filter simulating the acoustic echo path before being mixed [11], [12]. The simulation environment was designed to fit a small office room having different environmental conditions, as shown in Fig. 3 and Table I. It was found that the echo level measured at the microphone was 3.5-dB lower than that of the near-end speech on average. In particular, we substantially considered the changing echo path case (From II to VII in Table I) as well as the static echo path model (I in Table I). For simulating the varying echo environment, the echo path change was to occur at every 0.5-s interval. For simulating echo path variation, the room response changes every 0.5-s according to

\[
\text{II} \rightarrow \text{III} \rightarrow \text{IV} \rightarrow \text{V} \rightarrow \text{VI} \rightarrow \text{VII}
\]

in Table I. In order to create noisy conditions, white, babble, car, office, and street noises from the NOISEX-92 database were added to the clean speech data at 10, 20, and 30-dB SNR [13]. Based on this simulated data, GMM training with sixteen mixture components was performed with manual labeling to construct the classification models for single-talk and double-talk. We inherently made reference decisions on the clean speech material by hand-labelling at every frame.

For testing, we incorporated different 320-s long speech data to demonstrate the proposed GMM-based method is not biased to speech data used in training. Using the experimental conditions, the overall \(P_E\) results are shown in Tables II and III.

**TABLE I**

<table>
<thead>
<tr>
<th>MODELING ENVIRONMENT FOR IMPULSE RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room impulse response type</td>
</tr>
<tr>
<td>Size of space</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Location of speaker</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Location of microphone</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Reflection coefficient</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>COMPARISON OF (P_E) FOR ONE STATIC ECHO PATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise type</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Babble</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Car</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Office</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Street</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Average</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Table II, which indicates the results in the static echo path case, we can see that the $P_E$'s were significantly reduced by adopting the proposed method for all the noises and SNRs. We can observe that the proposed algorithm improved the performance up to 11.41% (overall average) compared to the time-domain approach. In addition, we can see in Table III, which shows the results for the case of echo path change, that the proposed algorithm gives us better performance than the previous approach in all environmental conditions (average difference = 14.05%).

A typical example of real-time classification results in conjunction with the original speech samples and manual marks is given in Figs. 4 (one static echo path) and 5 (rotating over six echo paths) for an easy understanding of the performance difference. From the figures, we can see that the proposed method effectively classifies the double-talk and single-talk, on the frame of the decision marks compared to those of the method of [7].

V. CONCLUSIONS

In this letter, we devised a novel method to improve the performance of the DTD. The proposed method is based on the GMM and utilizes a set of frequency-domain feature vectors, showing good performance in terms of the DTD accuracy. By incorporating the GMM-based DTD algorithm into the AES framework, robust reduction of the acoustic echo can be successfully achieved. Further improvement is expected if we find a more relevant feature for the double-talk situation.

TABLE III

<table>
<thead>
<tr>
<th>Noise type</th>
<th>Method</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>White</td>
<td>time-domain [7]</td>
<td>39.89±0.11</td>
</tr>
<tr>
<td>Babble</td>
<td>time-domain [7]</td>
<td>35.95±0.10</td>
</tr>
<tr>
<td>Car</td>
<td>time-domain [7]</td>
<td>34.83±0.10</td>
</tr>
<tr>
<td>Office</td>
<td>time-domain [7]</td>
<td>37.01±0.09</td>
</tr>
<tr>
<td>Street</td>
<td>time-domain [7]</td>
<td>33.84±0.11</td>
</tr>
<tr>
<td>Average</td>
<td>time-domain [7]</td>
<td>36.03±0.10</td>
</tr>
</tbody>
</table>

REFERENCES