A STATISTICAL MODEL-BASED DOUBLE-TALK DETECTION INCORPORATING SOFT DECISION

Yun-Sik Park, Ji-Hyun Song, Sang-Ick Kang, Wojuan Lee and Joon-Hyuk Chang

School of Electronic Engineering
Inha University, Incheon, Korea
E-mail: [yspark, jhsong, sikang, wjlee]@dsp.inha.ac.kr, changjh@inha.ac.kr

ABSTRACT

In this paper, we propose a novel double-talk detection (DTD) technique based on a soft decision in the frequency domain. The proposed method provides an efficient procedure to detect the double-talk situation by the use of the global near-end speech presence probability (GNSPP) and voice activity detection (VAD) of the near-end and far-end signal. Specifically, the GNSPP is derived based on a statistical method of speech and is employed to determine the double-talk presence in a given frame. The performance of our approach is evaluated by objective tests under different environments, and it is found that the suggested method yields better results compared with the conventional scheme.

Index Terms— Double-Talk Detection, Speech Presence Probability, Voice Activity Detection

1. INTRODUCTION

In most hands-free mobile communication systems, since the loudspeaker and microphone are acoustically coupled, acoustic echoes occur. In efforts to address this problem, numerous acoustic echo cancellation (AEC) techniques incorporating an adaptive filter such as the least mean square (LMS) and normalized LMS (NLMS) have been reported [1]-[3]. One of the major problems of AEC techniques, however, is that the performance significantly degrades during the double-talk periods, in which signals from both the near-end and far-end coexist because the double-talk acts as very large interference to the adaptive filter. The problem can be alleviated by freezing the adaptive filter coefficients through the use of a double talk detection (DTD) algorithm [4]. In this regard, many studies have been dedicated to the problem of DTD. In practice, cross-correlation and coherence-based approaches are relevant, as they present straightforward approaches. Adopting hard decisions [4]-[6], these schemes classify each frame into one of two (i.e., double-talk or not) cases by comparing decision statistics and given threshold values. However, they are sensitive to optimized parameters and do not always provide reliable performance under various conditions.

In this paper, we propose a novel DTD algorithm based on a global soft decision [7], where the term ‘global’ means that DTD is performed globally in a given frame and ‘soft decision’ [8], [9] denotes that the probability of double-talk is introduced as a decision and is applied to update the adaptive filter in the acoustic echo suppressor (AES) algorithm [10]. Specifically, the global near-end speech presence probability (GNSPP) based on a statistical model is computed in each frame to apply the proposed DTD algorithm in conjunction with results of voice activity detection (VAD) of the near-end and far-end signal. It is worth noting that our approach provides for the first time an effective framework of DTD based on a soft decision by taking advantage of a statistical model, in contrast with the conventional hard decision-based method. The performance of the proposed algorithm is evaluated by echo return loss enhancement (ERLE) and speech attenuation (SA) tests during double-talk and is demonstrated to be better than that of the conventional method.

2. GLOBAL NEAR-END SPEECH PRESENCE PROBABILITY

In this section, we consider how to derive the global near-end speech presence probability (GNSPP) in the frequency domain. To this end, we first assume that two hypotheses, $H_0$ and $H_1$, indicate near-end speech absence and presence as follows:

$$H_0 \ : \ \text{near-end speech absent} \ : \ Y(i) = D(i)$$
$$H_1 \ : \ \text{near-end speech present} \ : \ Y(i) = D(i) + S(i)$$

where $D(i) = [D(i,1), D(i,2), \ldots, D(i,M)]$, $S(i) = [S(i,1), S(i,2), \ldots, S(i,M)]$ and $Y(i) = [Y(i,1), Y(i,2), \ldots, Y(i,M)]$, respectively, represent the Fourier domain spectra of the echo signal, the near-end speech and the microphone input signal with a frame index $i$. Also, $X(i) = [X(i,1), X(i,2), \ldots, X(i,M)]$ denote the Fourier spectrum
of the far-end signal as shown in Fig. 1. The background noise is not taken into account since we assume that near-end speech absence is not correlated with the background noise. Under the assumption that \( D(i, k) \) and \( S(i, k) \) are characterized by separate zero-mean complex Gaussian distributions, the following is obtained [7]:

\[
p(Y(i, k) | H_0) = \frac{1}{\pi \lambda_d(i, k)} \exp \left[ -\frac{|Y(i, k)|^2}{\lambda_d(i, k)} \right] \tag{2}
\]

\[
p(Y(i, k) | H_1) = \frac{1}{\pi (\lambda_s(i, k) + \lambda_d(i, k))} \exp \left[ -\frac{|Y(i, k)|^2}{\lambda_s(i, k) + \lambda_d(i, k)} \right] \tag{3}
\]

where \( \lambda_s(i, k) \) and \( \lambda_d(i, k) \) are the variance of the near-end speech and estimated echo, respectively. Accordingly, the GNSPP \( p(H_1 | Y(i)) \) is derived from Bayes’ rule, such that [7]:

\[
p(H_1 | Y(i)) = \frac{p(Y(i) | H_1)p(H_1)}{p(Y(i) | H_0)p(H_0) + p(Y(i) | H_1)p(H_1)} \tag{4}
\]

where \( p(H_0)(= 1 - p(H_1)) \) represents the a priori probability of near-end speech absence. Since the spectral component in each frequency bin is assumed to be statistically independent, (4) can be rewritten as [7]

\[
p(H_1 | Y(i)) = \frac{p(H_1) \prod_{k=1}^{M} p(Y(i, k) | H_1)}{p(H_0) \prod_{k=1}^{M} p(Y(i, k) | H_0) + p(H_1) \prod_{k=1}^{M} p(Y(i, k) | H_1) + q \prod_{k=1}^{M} \Lambda_k(Y(i, k))} \tag{5}
\]

\[
= \frac{q \prod_{k=1}^{M} \Lambda_k(Y(i, k))}{1 + q \prod_{k=1}^{M} \Lambda_k(Y(i, k))}
\]

where \( q = p(H_1)/p(H_0)(= 1) \) which is determined by the rough estimate of the ratio of absence time duration and presence time duration for near-end speech and \( \Lambda_k(Y(i, k)) \) is the likelihood ratio computed in the \( k \)th frequency bin, as given by [7]:

\[
\Lambda_k(Y(i, k)) = \frac{p(Y(i, k) | H_1)}{p(Y(i, k) | H_0)} \tag{6}
\]

\[
= \frac{1}{1 + \xi(i, k)} \exp \left[ \frac{\gamma(i, k) \xi(i, k)}{1 + \xi(i, k)} \right]
\]

where the a posteriori signal-to-echo ratio (SER) \( \gamma(i, k) \) and the a priori SER \( \xi(i, k) \) are defined by

\[
\gamma(i, k) = \frac{|Y(i, k)|^2}{\lambda_d(i, k)} \tag{7}
\]

\[
\xi(i, k) \equiv \frac{\lambda_s(i, k)}{\lambda_d(i, k)} \tag{8}
\]

where \( \lambda_d(i, k) \) is estimated by \( \hat{\lambda}_d(i, k) \). The power spectrum of the echo signal is obtained in the case of the absence of the near-end speech signal, as given by

\[
\hat{\lambda}_d(i, k) = \xi_D \hat{\lambda}_d(i - 1, k) + (1 - \xi_D) |\hat{Y}(i, k)|^2 \tag{9}
\]

where \( \xi_D (= 0.93) \) is the smoothing parameter. Also, in (8), \( \xi(i, k) \) is estimated with the help of the well-known decision-directed approach with \( \alpha_{DD} = 0.6 \) [11]. Then,

\[
\hat{\gamma}(i, k) = \alpha_{DD} \frac{|\hat{S}(i - 1, k)|^2}{\lambda_d(i - 1, k)} + (1 - \alpha_{DD}) P\{\gamma(i, k) - 1\} \tag{10}
\]

where \( P\{z\} = z \text{ if } z \geq 0, \text{ and } P\{z\} = 0 \text{ otherwise. As specified in (10), the robust estimation of the echo magnitude spectrum } \hat{Y}(i, k) \text{ plays an essential role in the performance. In our approach, we follow the parameter estimation procedure proposed in [10] as follows:}

\[
|\hat{Y}(i, k)| = \hat{H}(i, k) |X(i, k)| \tag{11}
\]

where \( \hat{H}(i, k) \) is the estimate for the echo path response mimicking the actual echo path. Specifically, \( \hat{H}_{opt}(i, k) \) is obtained based on the magnitude of the least squares estimator as follows [10]:

\[
\hat{H}_{opt}(i, k) = \left[ \frac{E[X^∗(i, k)Y(i, k)]}{E[X^∗(i, k)X(i, k)]} \right] \tag{12}
\]

where \( * \) denotes the complex conjugate and \( E[\cdot] \) represents the expected value. Note that there exist some delay between the far-end speech \( X(i, k) \) and the microphone input signal \( Y(i, k) \) (due to a digital amplifier, e.g.). In our approach, it is assumed that the echo time-delay is separately estimated and compensated (i.e., no delay) at the near-end. Since the echo path is time varying, the estimated echo path response \( \hat{H}(i, k) \) is obtained using the iterative procedure such that[10]

\[
\hat{H}(i, k) = \frac{C(i, k)}{\hat{R}(i, k)} \tag{13}
\]

Fig. 1. Block diagram of the proposed DTD algorithm.
Table 1. ERLE and SA test results obtained from the proposed DTD algorithm based on a soft decision with those yielded by the conventional hard decision method during double-talk.

<table>
<thead>
<tr>
<th>Environments</th>
<th>ERLE (dB)</th>
<th>SA (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed Park-based</td>
<td>Proposed Park-based</td>
</tr>
<tr>
<td>Noise</td>
<td>SNR</td>
<td>Park-based</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>10</td>
<td>2.23</td>
</tr>
<tr>
<td>Babble</td>
<td>20</td>
<td>3.46</td>
</tr>
<tr>
<td>Vehicle</td>
<td>30</td>
<td>3.51</td>
</tr>
<tr>
<td>Clean speech</td>
<td>40</td>
<td>3.36</td>
</tr>
</tbody>
</table>

is known that it gives us a robust performance under various noise environments [12]. Furthermore, we modified the VAD algorithm to reduce the false decisions. For example, if the near-end signal $Y(i)$ exists at the $r$th frame and $I(Y(i)) = 0$ otherwise. Therefore, the update of $\hat{H}(i,k)$ is finally addressed such that $\hat{H}(i,k)$ replaces $\tilde{H}(i,k)$ (i.e., no update) within the double-talk regions on each frequency bin and (12) in the case of single-talk. In particular, in the case of abrupt transient periods between double-talk and single-talk, as shown in Fig. 2, the GNSPP could be a soft value between 0 and 1. This accounts for why the soft decision scheme is more insensitive to detection error compared to conventional hard decision methods.

Based on this proposed DTD method, we finally apply it to the AES algorithm proposed by Faller et al. [10] as follows:

\[
\tilde{S}(i,k) = G(i,k)Y(i,k)
\]

where the Wiener filter gain $G(i,k)$ is given by [10]

\[
G(i,k) = \left[ \frac{\max(|Y(i,k)| - |\hat{Y}(i,k)|, 0)}{|\hat{Y}(i,k)|} \right].
\]

4. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed DTD algorithm, we conducted objective comparison experiments under various noise conditions. Twenty test phrases, spoken by seven speakers and sampled at 8 kHz, were used as the experimental data. For assessing the performance of the proposed method, we artificially created 20 data files, where each file was produced by mixing the far-end signal with the near-end signal. Each frame of the windowed signal was transformed into its corresponding spectrum through a 128-point DFT after zero padding. We then constructed 16 frequency bands through combination of subbands to cover all frequency ranges (~4 kHz) of the narrow band speech signal,
which is analogous to that of the IS-127 noise suppression algorithm [12]. The far-end speech signal was passed through a filter simulating the acoustic echo path modeled by a time-invariant FIR filter based on the analysis of room acoustics before being mixed electrically [13], [14]. The simulation environment was designed to fit a small office room having a size of $5 \times 4 \times 3 \, m^3$. The echo level measured at the input microphone was 3.5 dB lower than that of the input near-end speech on average. In order to create noisy conditions, white, babble and vehicular noises from the NOISEX-92 database were added to clean near-end speech signals at signal-to-noise ratios (SNRs) of 10, 20, and 30 dB. For the purpose of an objective comparison, we evaluated the performance of the proposed scheme and that of the conventional DTD algorithm proposed by Park et al. [6], wherein the cross-correlation coefficients-based double-talk detection method is used. The performances of the approaches were measured in terms of echo return loss enhancement (ERLE) and speech attenuation (SA) during double-talk [14].

Given the three types of noise environments, the ERLE and SAs scores were averaged to give final mean score results, as presented in Table 1. From Table 1, it is evident that in most noisy conditions, the proposed DTD algorithm based on a soft decision yielded a lower ERLE compared to the hard decision-based conventional technique. Also, from Table 2, we can observe that the SAs (measured during double-talk periods) of the proposed scheme based on a soft decision were better than those of the previous method [6] in all the tested conditions. It is noted that the performance gain of the proposed method becomes smaller as the SNR becomes lower. This is attributed to imperfection of the GNSPP under adverse noise conditions. Summarizing the overall results, the proposed approach is found to be effective in the AES technique.

5. CONCLUSIONS

In this paper, we have proposed a novel DTD algorithm based on a soft decision scheme in the frequency domain. The GNSPP based on a statistical model of the near-end and far-end signal is applied to the DTD algorithm in conjunction with VAD decisions for effective echo suppression. The performance of the proposed algorithm has been found to be superior to that of the conventional technique through objective evaluation tests.

6. REFERENCES


