Objective and subjective evaluation of complementary Wiener filter for speech dereverberation

Kento Ohtani*, Tatsuya Komatsu, Kazunobu Kondo, Takanori Nishino and Kazuya Takeda

Acoustic distortion caused by reverberation can degrade speech quality and performance of speech-based systems. Several dereverberation techniques have been proposed in the literature. For example, a dereverberation method using a complementary Wiener filter can suppress late reverberation with few computational resources. As a method for dereverberation, the method using a complementary Wiener filter has been proposed, and for the exponential decay impulse response model, it is shown theoretically that we can suppress reverberation with few computational resources. In this report, we approximate expectation of the power spectrum which is necessary to calculate a complementary Wiener filter as exponential moving average. We conducted dereverberation experiments using actual environment room impulse response. The results of the objective evaluation show that the suppression performances of the actual environment room impulse response can approximate from the results of the exponential decay impulse response model. Additionally, we investigated the relationship between the results of objective evaluation and the results of subjective evaluation. In a small reverberation environment, we can see strong correlation between the results of objective and subjective evaluation.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Speech-based devices and systems, such as cellular phones and speech recognition software, are rapidly spreading. Accordingly, opportunities for the use of microphones are expanding as well. But these devices and systems are generally used under reverberant conditions such as in meeting rooms or automobiles, where speech quality and recognition performance are degraded by reverberation. As a result, several dereverberation techniques have been developed in recent years. Multi-microphone techniques have been utilized to estimate late reverberation based on a spatial correlation [1, 2, 3, 4, 5], or to estimate an inverse filter [6, 7] using the MINT theorem [8]. Multi-microphone techniques involve a large apparatus, and thus require the use of the single-channel method. But a serious drawback of the single-channel method is the lack of spatial information. Some single-channel speech enhancement techniques have shown successful results, such as the spectral subtraction method [9], MMSE-STSA [10], etc. These techniques have been successfully applied to methods of dereverberation [7, 11, 12, 13].

In a previous study, we proposed a single-channel dereverberation method using a complementary Wiener filter [14], and theoretically confirmed that our proposed method could suppress reverberation using limited computational resources under the exponential decay impulse response model. In this paper, we conduct dereverberation experiments using artificial and real reverberant speech, and approximate the expectation of the power spectrum, which is necessary to calculate a complementary Wiener filter, as an exponential moving average. Additionally, we conduct subjective evaluations of the dereverberated speech, and investigate the relationship between the results of the objective and subjective evaluations.

DEREVERBERATION METHOD

Dereverberated speech, based on a complementary Wiener filter, can be expressed as follows:

\[ Z[k, l] = G[k, l]X[k, l] \]  
(1)

where \( Z[k, l] \) represents dereverberated speech, \( X[k, l] \) represents the observed speech in the frequency domain, \( k \) represents the frequency bin index and \( l \) represents the frame index. \( G[k, l] \), which is the spectral gain of the dereverberation, is based on a complementary Wiener filter, which can be calculated as:

\[
G[k, l] = \begin{cases} 
1 & \frac{P_X[k, l]}{P_X[k]} \geq 1 \\
\frac{P_X[k, l]}{P_X[k]} & \text{(otherwise)} 
\end{cases}
\]  
(2)

where

\[
P_X[k, l] = |X[k, l]|^2
\]  
(3)
\[
P_X[k] = E_l[|X[k, l]|^2].
\]  
(4)

\( E_l[\cdot] \) is the expectation of the power spectrum over frame index \( l \). To solve the gain calculation, we approximate the expectation \( E_l[\cdot] \) using the exponential moving average, as shown in Eq. 5:

\[
R[k, l] = (1 - \alpha)|X[k, l]|^2 + \alpha R[k, l - 1].
\]  
(5)

\( \alpha (0 < \alpha < 1) \) in this equation is a weighting factor, which represents how many past values we consider, and is called a smoothing constant.
<table>
<thead>
<tr>
<th>Reverberant environment</th>
<th>Reverberation time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting room</td>
<td>0.361</td>
</tr>
<tr>
<td>Practice room</td>
<td>0.520</td>
</tr>
<tr>
<td>Recording booth</td>
<td>0.540</td>
</tr>
<tr>
<td>Live music club</td>
<td>0.677</td>
</tr>
<tr>
<td>Training room</td>
<td>0.864</td>
</tr>
<tr>
<td>Classroom</td>
<td>0.928</td>
</tr>
<tr>
<td>Meeting place</td>
<td>0.961</td>
</tr>
<tr>
<td>Auditorium</td>
<td>1.032</td>
</tr>
</tbody>
</table>

**OBJECTIVE EVALUATION**

We objectively evaluated dereverberation performance of the complementary Wiener filter. We then created target speech for evaluation by adding reverberation. We did this by convolving clean speech with room impulse response (RIR). We used simulated RIRs based on an exponential decay model and RIRs measured in real environments. Dereverberation performance using these RIRs was then compared.

**Evaluated room impulse responses**

To make artificial RIRs, we use the exponential decay model [15] as follows:

\[
h[n] = \begin{cases} 
  b[n]e^{-\Delta n} & (n \geq 0) \\
  0 & \text{(otherwise)} 
\end{cases} \tag{6}
\]

\[
\Delta = \frac{3\ln10}{T_r f_s} \tag{7}
\]

where \( b[n] \) is stationary white noise with zero mean, \( T_r \) is reverberation time and \( f_s \) is the sampling rate. Considering the randomness of \( b[n] \), we created white noise samples 100 times. For each white noise sample, we changed the reverberation time from 0.3s to 1.2s and created additional RIRs with different reverberation times.

For the real environment RIRs, we used the RIRs in an architectural acoustics database [16]. A list of the real environment RIRs used is shown in Table 1.

**Experiment**

We then evaluate the results using improvement in segmental target-to-interference ratio (TIR). TIR can be derived by averaging the ratio of target speech to interference speech in all frames as follows:

\[
\text{TIR[dB]} = 10 \log \frac{\sum_{n} |t[n]|^2}{\sum_{n} |x[n] - t[n]|^2} \tag{8}
\]

where \( t[n] \) represents the target speech, \( x[n] \) represents the speech being evaluated, \( l \) represents the frame index, and \( N \) represents the number of frames. A short RIR, tens of msec, is sufficient in length to measure speech clarity, and D50 and C80 were used as indexes [17]. Therefore, we used target speech which was convolved with 32 ms of RIR from the beginning of the sample.

\[
t[n] = \sum_{k=0}^{511} h[k]s[n-k] \tag{9}
\]
A larger TIR means fewer interference components (late reverberation components).

Experimental conditions are shown in Table 2. Smoothing constant $\alpha$ is changed according to reverberation time to maximize TIR improvement.

### Experimental results

Fig. 1 shows the dereverberation results for each reverberated speech sample. The horizontal axis represents reverberation time and the vertical axis represents TIR improvement. The curved line in the figure represents the result of the artificial RIR and the white dots represent the results of the real environment RIR. Each curved line shows a diversified reverberation time, obtained by diversifying $T_r$ in Eq. 7 to each white noise sample. We can see that as reverberation time increases, TIR improvement increases. If we conduct a similar experiment with another white noise sample, the shape of the curved line is not changed, but the position goes up and down. When we create white noise samples 100 times, TIR improvement varies within about 1 dB. We considered these results to be affected by the randomness of the white noise samples. The results of the real environment RIRs (marked by white dots) fall within the range of the results of the artificial environment, thus we can see that dereverberation performance in a real environment can be estimated from dereverberation performance of an artificial RIR.

In Fig. 2, the horizontal axis represents the smoothing constant and the vertical axis represents TIR improvement. As the smoothing constant increases, TIR improvement increases at first, but after maximum TIR improvement is reached, it then decreases gradually. Additionally, we can see that as reverberation time increases, the smoothing constant of maximum TIR improvement also increases.

### SUBJECTIVE EVALUATION

We are unable to evaluate the quality of speech by objective evaluation as we do musical noise, therefore, we conducted the subjective evaluations of some of the dereverberated speech...
and evaluated its acoustic speech quality. Additionally, we investigated the relationship between the results of objective evaluation and subjective evaluation. We created target speech for this evaluation in the same manner as we did for our objective evaluation; by convolving clean speech with RIR.

**Target speech for evaluation**

Evaluated RIRs are shown in Table 3. We prepared three reverberation times (around 0.35 s, 0.65 s, 1.00s), and for each reverberation time we prepared two types of real environment RIRs and one type of artificial RIR. In a manner similar to that used in the objective evaluation, for the real environment RIRs we used RIRs from the architectural acoustics database, and for the artificial RIRs we used the exponential decay model shown in Eqs. 6 and 7.

**Experiment**

Dereverberation conditions are shown in Table 4. From the results of the preceding section, we can see the relationship between the smoothing constant and TIR improvement. Therefore, we chose 5 smoothing constants; maximum of TIR improvement, 50% and 75% of maximum TIR improvement (Fig. 3). The number of test subjects was 8, and 4 of them had taken part in the preliminary experiment. Each of the test subjects evaluate 90 speech samples chosen in section. 4.1 (two persons - one male and one female, nine reverberant environments, and 5 smoothing constants). Speech samples were transduced using headphones (Audio-technica, ATH-A900). Each test subject could change the volume, and the order of the speech samples were random for each test subject.
TABLE 4: Dereverberation conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>16,000 Hz</td>
</tr>
<tr>
<td>Analysis window</td>
<td>Hanning</td>
</tr>
<tr>
<td>Frame length</td>
<td>32 ms (512 pt)</td>
</tr>
<tr>
<td>Shift width</td>
<td>8 ms (128 pt)</td>
</tr>
<tr>
<td>Smooth constant for each RIR</td>
<td>5 points</td>
</tr>
</tbody>
</table>

FIGURE 3: Choice of smoothing constants

Test subjects evaluated acoustic distortion of each speech sample according to a five point system (1: no acoustic distortion, 5: most distorted speech). We explained the reverberation time change to the subjects, and demonstrated standards for the five grades. To demonstrate speech with no acoustic distortion, we used the target speech in Eq. 9, and we choose the most distorted speech ourselves.

**Experimental results**

In Fig. 4, the horizontal axis represents the smoothing constant and the vertical axis represents TIR improvement. We can see that the average score is highest when the reverberation time is around 0.35 s, and that as reverberation time increases, the average score decreases. Especially when the reverberation time is around 1.00 s, we can see that dereverberation with no acoustic distortion is difficult, because the score is very low. Additionally, we can see the tendency that as the smoothing constant increases, the average score decreases.

Next, in Fig.5 and 6, we compare average distortion scores given by male and female
subjects, respectively, and TIR improvement. Both TIR improvement and average evaluation scores are higher when the results are good, (speech with less reverberation and less distortion will be plotted in upper right area of the graph). If we calculate the correlation coefficient for each reverberation time, when reverberation time is around 0.35 s, the correlation coefficients are 0.5642 (male) and 0.6263 (female). From this result we can see that, in environments with short reverberation, there is a strong correlation between the results of objective and subjective evaluation, while with higher reverberation times the correlation coefficient is lower.

**CONCLUSION**

We conducted objective and subjective evaluations of speech dereverberation using a complementary Wiener filter. The results showed that objective evaluation of artificial RIRs went up and down because of the randomness of white noise, but we could see that objective evaluations corresponded with the results of real environment RIRs. From these results, we can see that we can estimate dereverberation performance in real environments from the dereverberation performance of artificial RIRs.

From the results of the subjective evaluations, we can see that there is a strong correlation between the results of the objective and subjective evaluations in environments with short reverberation. Additionally, it is clear that as reverberation time increases, acoustic distortion increases as well.

Future work will investigate a method of estimating the smoothing constant for maximum TIR improvement.
ACKNOWLEDGMENTS

This work was partially supported by the Core Research for Evolutional Science and Technology (CREST) Program of the Japan Science and Technology Agency (JST).

REFERENCES


