Development of the MTF-based speech dereverberation method using adaptive time-frequency division

Masashi Unoki, Masato Toi, and Masato Akagi
School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 JAPAN e-mail: {unoki, m-toi, akagi}@jaist.ac.jp

We previously proposed a speech dereverberation method based on the modulation transfer function (MTF) concept. In that model, power envelopes and carriers are decomposed from a reverberant speech signal using a constant-bandwidth filterbank and then are restored in each respective channel using, respectively, power envelope inverse filtering and a carrier regeneration method. In this paper, we explain how the model can be improved through adaptive time-frequency division using a reconstructed filterbank that depends on each speech signal. In addition, the “buzz” sounds caused by the carrier regeneration process of our previous model can be removed by controlling the group delay in the carrier. We have carried out 15,000 simulations of dereverberation for reverberant speech signals to objectively evaluate the improved model. We also subjectively evaluated the model via listening tests (sound quality, reverberation, and word intelligibility). Results of both evaluations show that in addition to reducing the averaged log-spectrum distortion by about 1 dB the improved model reduces the loss of speech intelligibility caused by reverberation by 30%.

1 Introduction

In general, reverberation smears significant features in speech and lowers speech intelligibility. Speech dereverberation is therefore needed in various forms of speech signal processing, such as in hearing aid systems and the preprocessing for speech recognition systems. The ultimate goal of our work is to construct a blind speech dereverberation method which can restore a speech signal from reverberant speech without using useful prior information such as the impulse response of the room acoustics, and which enables less loss of speech intelligibility caused by reverberation.

Several well-known inverse filtering methods can be used to dereverberate the original signal from a reverberant signal in room acoustics. There are, for example, the methods of Neely and Allen [1] and Miyoshi and Kaneda [2]. These methods use a single microphone or microphone array to dereverberate signals. Miyoshi and Kaneda’s method can be applied to room acoustics with non-minimum phase characteristics, but Neely and Allen’s method can be applied only to the minimum phase characteristics in the room acoustics. For these methods, however, the impulse response of the room acoustics must be measured before the dereverberation to determine the inverse filtering. Moreover, the impulse response, in general, is time variant. These methods would therefore require remeasurement of the impulse response of the room acoustics for each dereverberation to be processed. This is a significant drawback with regard to the use of these methods for various speech applications.

Recently, Nakatani and Miyoshi proposed a blind dereverberation method for a single-channel speech signal based on harmonic structure without measuring the impulse response of the room acoustics [3]. This method, however, requires accurate estimation of the fundamental frequency from the reverberant speech, and they pointed out that it is difficult to meet this requirement. It also seems that this method does not precisely dereverberate the parts corresponding to consonants.

On the other hand, temporal envelope inverse filtering methods have been proposed that not only restore the temporal envelope from reverberant speech, but also improve speech intelligibility that is degraded by reverberation. There are, for example, the methods of Avendano and Hermansky [4] and Hirobayashi et al. [5]. These methods are based on the modulation transfer function (MTF) concept [6] and can restore the temporal envelope information (temporal fluctuation or modulation index) from reverberant speech without requiring measurement of the room’s impulse response. They will therefore be useful for preprocessing in such applications. However, they do not effectively improve the speech intelligibility of the restored speech signal. We think this shortcoming is due to artifacts in the fine-structure of the restored speech since these methods disregard the carrier of the speech signal in their dereverberation processing.

In our previous work, as the first step, we improved the model of Hirobayashi et al. [5] to solve the problems they encountered [7] and then extended the improved basic model to a model based on a filterbank by considering the issues regarding speech applications [8]. As the second step, we refined our filterbank model to enable restoration of a reverberant speech signal and reconstruction of a speech waveform after dereverberation [9]. In this paper, to develop an MTF-based speech dereverberation method that will enable recovery of the loss of speech intelligibility, we extend our model to allow adaptive time-frequency division processing and improve the carrier regeneration process in the model.

2 Speech dereverberation method

Figure 1 shows a block-diagram of our speech dereverberation model. The model consists of two main parts:
the power envelope restoration and the carrier regeneration processes [9]. In this work, we define $x(t)$, $y(t) = h(t) \ast x(t)$, and $h(t)$ as the original, the reverberant signal, and the stochastic-idealized impulse response of the room acoustics, respectively. In particular, $h(t)$ is defined as

$$h(t) = a \exp(-6.91/T_R) n(t) = e_h(t) n(t),$$  \hspace{1cm} (1)

where $a$, $T_R$, and $n(t)$ are a constant amplitude term, reverberation time, and random white noise, respectively.

First, the reverberant signal $y(t)$ is decomposed into envelopes $e_{y,n}(t)$ and carriers $c_{y,n}(t)$ in $N$-channels ($1 \leq n \leq N$) using an analysis filterbank. Here, it is assumed that $x(t)$ and $y(t)$ can be represented as

$$x(t) := \sum_{n=1}^{N} x_n(t) = \sum_{n=1}^{N} e_{x,n}(t) \cdot c_{x,n}(t),$$  \hspace{1cm} (2)

$$y(t) := \sum_{n=1}^{N} y_n(t) = \sum_{n=1}^{N} e_{y,n}(t) \cdot c_{y,n}(t),$$  \hspace{1cm} (3)

where $x_n(t)$ and $y_n(t)$ are the band-limited signals, $e_{x,n}(t)$ and $e_{y,n}(t)$ are the temporal envelopes, and $c_{x,n}(t)$ and $c_{y,n}(t)$ are the carriers in the channels, respectively. Second, in the power envelope dereverberation process, the power envelope inverse filtering method [7] is used to recover $\hat{e}_{x,n}(t)^2$ from $e_{y,n}(t)^2$. Then, in the carrier regeneration process, a source generator is used to reconstruct the source information of the dereverberated speech from voiced/unvoiced information based on the estimated fundamental frequency (F0), and then the source information is decomposed into the carriers in the channels, $\hat{c}_{x,n}(t)$. Finally, each channel signal is reproduced by multiplying the dereverberated envelope with the regenerated carrier (i.e., $\hat{x}_n(t) = \hat{e}_{x,n}(t) \cdot \hat{c}_{x,n}(t)$) in the corresponding channel, and the dereverberated signal $\hat{x}(t)$ is resynthesized using a synthesis filterbank, which is the same as the analysis filterbank.

### 2.1 Power envelope restoration process

In the power envelope restoration process, $e_{x,n}(t)^2$ can be restored from $e_{y,n}(t)^2$ using the relation $e_{y,n}(t)^2 = e_{x,n}(t)^2 \ast e_{h}(t)^2$ derived from the MTF concept (for details, see [7, 8]). Here, the transmission functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to be the respective $z$-transforms of $e_{x,n}(t)^2$, $e_h(t)^2$, and $e_{y,n}(t)^2$. Thus, $E_x(z)$, can be determined as

$$E_x(z) = \frac{E_y(z)}{a^N} \left\{ 1 - \exp \left( -\frac{13.8}{T_R \cdot f_s} \right) z^{-1} \right\},$$  \hspace{1cm} (4)

where $f_s$ is the sampling frequency (20 kHz). Here, to determine the parameters ($a$ and $T_R$), we apply the following algorithm [9] in each channel, separately.

$$\hat{e}_{y,n}(t)^2 = \text{LPF} \left( |y_n(t) + j \text{Hilbert}(y_n(t))|^2 \right),$$  \hspace{1cm} (5)

$$\hat{T}_R = \arg \min_{0 \leq T_R \leq T_{R\text{max}}} \left\{ \int_0^{T_R} |\hat{e}_{x,n}(t)^2 - \theta| dt \right\},$$  \hspace{1cm} (6)

$$T_P(T_R) = \min_{\theta_{\text{min}} \leq \theta \leq \theta_{\text{max}}} \left( |\hat{e}_{x,n}(t)^2 - \theta| \right),$$  \hspace{1cm} (7)

$$\hat{a} = \sqrt{1/\int_0^{T} \exp(-13.8t/T_R) dt},$$  \hspace{1cm} (8)

where Hilbert() is the Hilbert transform, LPF[] is low-pass filtering with a cut-off frequency of 20 Hz, $e_{x,n}(t)^2$ is the restored power envelope using any $T_R$ within the given conditions, $\theta$ is a threshold for detecting a point from the maximum of $e_{y,n}(t)^2$, and $\theta_{\text{min}}$ and $\theta_{\text{max}}$ are the lower and the upper limited regions for determining the point. Finally, $e_{x,n}(t)^2$ can be obtained from the inverse $z$-transform of $E_x(z)$ for each channel.

### 2.2 Carrier regeneration process

Figure 2 shows the signal flow of the carrier regeneration process. This processing is done separately for the voiced and unvoiced intervals, which are estimated using the estimated fundamental frequency (F0). In this paper, we assume that F0 has been estimated accurately.

First, the carrier regeneration in the voiced intervals is based on some known facts: the source information of voiced speech consists of harmonicity and/or periodicity with the fundamental frequency (F0) and this harmonicity can smoothly vary with the F0 fluctuation; in addition, the harmonicity is frequency-modulated together with F0. Thus, harmonicity with the estimated $F_0(t)$ can be regenerated based on the pulse source interpolated frequency modulation (PIMF) model [10]. The regenerated carrier $\hat{c}_v(t)$ in the voiced intervals is represented as

$$\hat{c}_v(t) = \frac{1}{K} \sum_{k=1}^{K} \sin \left( k \int_0^t 2\pi F_0(\tau)d\tau + \phi_k(t) \right),$$  \hspace{1cm} (9)

where $F_0(t)$ is the fundamental frequency, $\phi_k(t)$ is the initial phase, $k$ is the index of harmonics ($1 \leq k \leq K(t)$), and $K(t)$ is the maximum number of harmonics.

Next, the carrier regeneration in the unvoiced intervals regenerates a white noise carrier $c_w(t)$, instead of $c_v(t)$. 

---

[Figure 1: MTF-based speech dereverberation model.]

---

[Figure 2: Signal flow of the carrier regeneration process.]
All regenerated carriers are then added together; i.e., \( \hat{c}_x(t) = \hat{c}_v(t) + \hat{c}_u(t) \). The resultant carrier \( \hat{c}_v(t) \) is then decomposed into \( \hat{c}_{v,n}(t) \) in separate channels using the analysis filterbank where the amplitude is normalized to 1 by the power envelope. Here, we did not deal with \( \phi(t) \) so we set \( \phi(t) = 0 \).

3 Improved method

Figure 3 shows an example of a dereverberation result when we used the proposed model for a Japanese sentence (/aikawarazu/) of a male speaker (Mau) from the ATR speech database [12] and a reverberation time of \( T_R = 1.0 \) s. This figure shows the amplitude spectrograms for (a) the original \( x(t) \), (b) reverberant signal \( y(t) \), and (c) the dereverberated speech \( \hat{x}(t) \) obtained using the previous model. The reverberation causes irregularity that destroys harmonicity and the relationship between peaks and dips in the time-frequency region, as can be seen when we compare (a) and (b). This figure, when we compare (a) and (c), shows that the previous model can recover harmonicity and a silence interval from the reverberant spectrum.

In the proposed model, however, the constant-band \( N \)-channel filterbank \( (N = 100, \text{ bandwidth of } 100 \text{ Hz}) \) is used to decompose the power envelopes and the carriers from the signal. The power envelope restoration process is done for a whole range in each channel. So, it may be possible to improve the restoration accuracy through adequate time and frequency division in the model. In addition, the carrier regeneration process is done without manipulating the phase of the carrier, which causes artifacts (such as buzz sounds) that can be heard and reduce sound quality.

We thus reconsidered our previous model with regard to two points: (1) how to design a filterbank (to enable time division and frequency division) to best improve the restoration accuracy in the power envelope restoration model; and (2) how to develop a carrier reconstruction method that improves the sound quality and the intelligibility of the restored sound. In this section, we explain how we have improved the MTF-based speech dereverberation model with regard to these points.

3.1 Adaptive time-frequency division

With regard to frequency division, the co-modulation characteristics of the target speech are used to determine the appropriate bandwidth in the filterbank. We examined the correlation between the power envelopes in the channels with a constant narrow-band (40 Hz) filterbank to verify the co-modulation characteristics. Figure 4, for example, shows a contour which is the region of correlation over 0.98 representing the co-modulation characteristics. This contour tended to widen as the region of correlation decreased. From this contour, we regarded a bandwidth of 90 Hz as an approximate constant bandwidth.
In this paper, we define parameter $r$ as the boundary of correlation for the co-modulation characteristics to determine an adequate bandwidth. Adequate bandwidths are determined from contour-widths from lower to higher frequencies (as a contour-width as shown in the right panel of Fig. 4).

With regard to time division, we use a threshold method to determine time segments from a whole duration in each channel. In this paper, we assume that each center point within $\eta$ dB down from the maximum value in $e_{x,n}(t)^2$ shows the boundary for segments in each channel. Figure 5 shows, for example, the power envelopes on the 76th-channel ($n = 76$) in Fig. 3 when $\eta = -10$ dB. The thick solid line shows an original power envelope $e_{x,n}(t)^2$. The thinner solid line and the dashed line show restored and reverberant power envelopes, $e_{x,n}(t)^2$ and $e_{y,n}(t)^2$, respectively. Figure 5 (a) shows the model case, while Fig. 5 (b) shows three segments for restoration in the same channel when using the threshold method. $T_R$ was 0.56, 1.50, and 0.41 s in the three segments while $T_R$ was 0.29 s in a whole duration. Comparing the matched shape between $e_{x}(t)^2$ and $e_{y}(t)^2$ in the two panels, we can see that $e_{x}(t)^2$ in the time-division case was restored more accurately than with the threshold method.

### 3.2 Carrier regeneration process with group delay control

With our previous process (Fig. 2), the phase information $\phi(t)$ in the carrier is disregarded and the sound quality of the restored speech is degraded the presence of a buzz sound. To remove artifacts such as buzz sound, we tried to directly manipulate $\phi(t)$ (e.g., by randomizing the initial $\phi(t)$ or fitting $\phi(t)$ to the data); however we found no effective way to do this.

For our current model, we applied the group delay control used in STRAIGHT [11] to indirectly manipulate $\phi(t)$ in the carrier regeneration process. Thus, we designed an all-pass filter which has units for the amplitude spectrum and the random phase spectrum, where the group delay $\tau_g$ is defined as

$$\tau_g(\omega) = \frac{\rho(\omega) - d_g v(\omega)}{\sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} |v(\omega)|^2 d\omega}}$$

where $N(\tau)$ is the Fourier transform $\mathcal{F}$ of random noise, $\mathcal{F}^{-1}$ is the inverse Fourier transform, $W_s(\tau)$ is the weighting function, and $\rho(\omega)$ is a weighting on the frequency $\omega$. We controlled this all-pass filter through the group delay $\tau_g$, in which $b_w = 3$ kHz and $d_g = 4$ ms, and then convolved it with the voiced source carrier $c_v(t)$.

### 4 Evaluation

#### 4.1 Simulation

We carried out simulations to evaluate the improved model with regard to time and/or frequency division, comparing it with our previous model. The speech signals were three Japanese sentences (/ai/kawarazu/, /shin-bun/, and /foudan/) uttered by ten speakers (five males and five females) from the ATR-database [12]. We used 100 types of impulse response $h(t)$, and five reverberation times $T_R = 0.1, 0.3, 0.5, 1.0,$ and $2.0$ s. All stimuli, $y(t)$ were composed through $15,000 (= 3 \times 10 \times 5 \times 10)$ convolutions of $x(t)$ with $h(t)$.

In these simulations, correlation coefficient of $(e_x^2, e_y^2)$ ($\text{Corr}(e_x^2, e_y^2)$, see also [8]), SNR($e_x^2, e_y^2$), and log-spectrum distortion (LSD) were used as evaluation measure are used to show the improvement in restoration with our improved model [8, 9].

$$\text{SNR}(e_x^2, e_y^2) = 20 \log_{10} \frac{\int_{0}^{T} e_x^2(t)^2 dt}{\int_{0}^{T} (e_x^2(t)^2 - e_y^2)^2 dt}, \quad \text{(dB)}$$

$$\text{LSD} = \left( \frac{1}{W} \sum_{\omega} \left( 20 \log_{10} \left| \frac{S(\omega)}{|\hat{S}(\omega)|} \right| \right)^2 \right), \quad \text{(dB)}$$

where $W$ is the upper frequency (here, it was 10 kHz), and $|S(\omega)|$ and $|\hat{S}(\omega)|$ are the amplitude spectra of $\hat{x}(t)$ and $\hat{y}(t)$ or $\hat{x}(t)$, respectively. Here, $\hat{x}(t)$ is the original speech resynthesized using the carrier generation, and $\hat{y}(t)$ is the reverberant $\hat{x}(t)$ (i.e., $\hat{x}(t) \ast h(t)$). Figure 3...
Figure 8 shows that while all four models reduced the per, we refer to the fourth model as the improved model. We then compared four models (previous, time-division, frequency-division, and time-frequency division process- ing) to determine the improvement attainable through use of time- and/or frequency-division processing. In this paper, we refer to the fourth model as the improved model. Figure 8 shows that while all four models reduced the accuracy per Hertz where $\eta$ obtained in frequency division with $r$ (frequency division) to determine the improvement attainable through use of time-frequency division. These figures show that the greatest improvement can be achieved about 30% of the reduction in speech intelligibility caused by the reverberation. We thus believe that the improved speech quality greatly contributed to improved speech intelligibility.

In these evaluations, we first determined whether we could obtain adequate $r$ in frequency division and adequate $\eta$ in time division for all results, based on the improved correlation and SNR. Figure 6 shows the averaged improvement in restoration accuracy (correlation and SNR) per Hertz for the power envelope in frequency division where $r = 0.70, 0.8, 0.95, 0.98, \text{and} 0.99.$ Figure 7 shows the averaged improvement in restoration accuracy per Hertz where $\eta = 20, 15, 10, \text{and} -8 \text{\, dB}$. These figures show that the greatest improvement can be obtained in frequency division with $r = 0.98$ or in time division with $\eta = -10 \text{\, dB}$. In this paper, we used these values to achieve adequate time-frequency division.

We then compared four models (previous, time-division, frequency-division, and time-frequency division processing) to determine the improvement attainable through use of time- and/or frequency-division processing. In this paper, we refer to the fourth model as the improved model. Figure 8 shows that while all four models reduced the

<table>
<thead>
<tr>
<th>Threshold, $\eta$ (dB)</th>
<th>Mean of Improved Correlation</th>
<th>Improved LSD (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
<td>1.98</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>1.95</td>
</tr>
<tr>
<td>6</td>
<td>0.35</td>
<td>1.92</td>
</tr>
<tr>
<td>8</td>
<td>0.45</td>
<td>1.89</td>
</tr>
<tr>
<td>10</td>
<td>0.55</td>
<td>1.86</td>
</tr>
</tbody>
</table>

Figure 8: Improvement in restoration accuracy (LSD).

LSD, the restoration was better when using time division and/or frequency division with the previous model. The improved method provided the greatest improvement in restoration accuracy under all conditions: LSD was about 1 dB lower than that with the previous model.

4.2 Listening tests

We also evaluated our previous and current models through listening tests. We conducted the listening tests in a soundproof room with five subjects who had normal hearing. We used Scheffe’s method of paired comparison (“Which is better, for A or B?”) to measure the sound quality (not including a reverberant speech) and the reverberation of restored speech in 5 graded evaluations ($-2, -1, 0, 1, \text{and} 2$) as perceived by the subjects. In these experiments, the speech signals were eight words of Japanese sentences from an NTT database for speech intelligibility testing [13], and the stimuli were manipulated under four conditions (original, reverberant, and dereverberated by the previous and proposed models).

Figure 9(a) shows the sound quality and Fig. 9(b) shows the reverberation. The sound quality of the improved model was better than that of the previous model. However, subjects experienced more reverberation with the improved model than with the previous model. By adjusting the group delay, though, we will be able to reach a compromise between sound quality and reverberation where the dereverberated speech sounds more natural.

We also measured the speech intelligibility of the dereverberated speech with the same subjects. The speech signals were 12 words (4 familiarity levels × 3 words) from the database. In this test, subjects were asked to compare the intelligibility of the four versions (reverberant, dereverberated by the previous model, dereverberated by the improved model, and original) of each word. Word intelligibility improved as the familiarity level increased, and was higher with the improved model than with the previous model or for the reverberant version (Fig. 10). In particular, the improved model recovered about 30% of the reduction in speech intelligibility caused by the reverberation. We thus believe that the improved speech quality greatly contributed to improved speech intelligibility.
5 Conclusion

We have improved our previous speech dereverberation method based on the MTF concept, to extend it for adaptive time-frequency division processing. This was done through use of a reconstruction filterbank that depends on each speech signal and refinement of the carrier regeneration process by adjusting the group delay in the carrier to removed artifacts such as buzz sounds.

To evaluate the model incorporating with these methods, we carried out objective and subjective experiments. Through 15,000 simulations of the dereverberation of reverberant speech, we found that the greatest improvement in restoration accuracy for the power envelope can be obtained through frequency division with $r = 0.98$ or time division with $\eta = -10$ dB. We also found that the improved model could adequately restore reverberant speech and provided the best restoration accuracy of the four methods. The LSD was improved by about 1 dB compared with that of the previous model. The results of three subjective experiments showed that the model could accurately dereverberate the reverberant speech and reduce the loss of speech intelligibility caused by reverberation. Envelope restoration based on the MTF concept and carrier reconstruction related to group delay control for speech synthesis contributed to the recovery of reverberant speech with regard to the signal and speech intelligibility.

In our future work, we will attempt to develop a method to accurately estimate F0 from reverberant speech and will investigate the dereverberation accuracy of the proposed model in real-life situations.

6 Acknowledgment

This work was supported by a Grant-in-Aid for Science Research from the Ministry of Education (No. 14780267, and 17650048).

References