A speech dereverberation method based on the MTF concept

Masashi Unoki, Keigo Sakata, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292 JAPAN
{unoki,k-sakata,akagi}@jaist.ac.jp

Abstract
This paper proposes a speech dereverberation method based on the MTF concept. This method can be used without measuring the impulse response of room acoustics. In the model, the power envelopes and carriers are decomposed from a reverberant speech signal using an N-channel filterbank and then are dereverberated in each respective channel. In the envelope dereverberation process, a power envelope inverse filtering method is used to dereverberate the envelopes. In the carrier regeneration process, a carrier generation method based on voiced/unvoiced speech from the estimated fundamental frequency (F0) is used. In this paper, we assume that F0 has been estimated accurately. We have carried out 15,000 simulations of dereverberation for reverberant speech signals to evaluate the proposed model. We found that the proposed model can accurately dereverberate not only the power envelopes but also the speech signal from the reverberant speech using regenerated carriers.

1. Introduction
Recovery of the original speech signal from a reverberant speech signal is an important issue concerning not only various kinds of speech signal processing such as speech emphasis for transmission systems (speaker to microphone) and hearing aid systems, but also regarding preprocessing for robust speech recognition systems.

Inverse filtering methods have been proposed to dereverberate the original signal from the reverberant signal in room acoustics. For example, Neely and Allen proposed a method that used a single microphone to remove a minimum phase component from the room effect [1]. This method, however, can only be used for room acoustics with minimum phase characteristics. Miyoshi and Kaneda proposed another method that used a microphone array and constrained non-overlaps of zeros in all pairs of the impulse responses between the sources and the microphones [2]. This method can be applied to room acoustics with non-minimum phase characteristics. However, these methods require that the impulse response of the room be measured to determine the inverse filtering before the dereverberation. Moreover, the impulse response temporally varies with various environmental factors (temperature, etc.), so the room acoustics have to be measured each time these methods are used. This is a significant drawback with regard to these methods.

On the other hand, Hirobayashi et al. proposed the power envelope inverse filtering method [3]. This method, based on the modulation transfer function (MTF) [4], can be used to recover the power envelope of the original signal from the reverberant signal without measuring the impulse response of the room. In addition, we have proposed an improved method [5] that solves problems with their method: how to precisely extract the power envelope from the observed signal and how to determine the parameters of the reverberant time and the amplitude terms ($T_R$ and $a$) of the impulse response. However, there are certain problems with this method concerning speech applications, as described in Sec. 3.2. Moreover, this type of method can be applied only to preprocessing for speech recognition systems even if we can solve these problems, though this method cannot reconstruct speech waveform after dereverberation, for instance, to apply to hearing aid systems.

In this paper, we propose a speech dereverberation method, based on the MTF concept, for dereverberating the power envelopes and the carriers of the filterbank from the reverberant speech.

2. Speech dereverberation method
The proposed speech dereverberation model is shown in Fig. 1. This model consists of two parts: the power envelope dereverberation and the carrier regeneration. First, the reverberant speech signal is decomposed into the envelopes and the carriers in N-channels (bandpass filtering) using a filterbank (analysis processing). Then, in the envelope dereverberation process, the power envelope inverse filtering method is used to recover the power envelope of speech from the reverberant envelope, as described in Sec. 3. In the carrier regeneration process, a source generator based on the pulse source interpolated frequency modulation (PIFM) method [6] is used to reconstruct the source information of the dereverberated speech from voiced/unvoiced information based on the estimated fundamental frequency, and then the source information is decomposed into the carriers in channels, as described in Sec. 4. Finally, each channel signal is reproduced by multiplying the dereverberated envelope with the regenerated carrier, and then the channel signals are reconstructed into the dereverberated signal using the same filterbank (resynthesis processing).
3. Power envelope inverse filtering

3.1. Model concept based on the MTF

In a basic model for inverse-filtering of the power envelope proposed by Hirobayashi et al. [3], the observed reverberant signal, the original signal, and the stochastical-idealized impulse response in the room acoustics [4] are denoted as $y(t)$, $x(t)$, and $h(t)$, respectively, and are modeled as follows:

$$y(t) = x(t) * h(t),$$  \hspace{1cm} (1)

$$x(t) = e_x(t) n_1(t),$$  \hspace{1cm} (2)

$$h(t) = e_h(t) n_2(t) = a \exp(-6.9 t/T_R) n_2(t),$$  \hspace{1cm} (3)

where \(a\) denotes the convolution operation, $e_x(t)$ and $e_h(t)$ are the envelopes of $x(t)$ and $h(t)$, and $n_1(t)$ and $n_2(t)$ are the mutually independent respective white noise functions, i.e., $\langle n_1(t), n_2(t - \tau) \rangle = \delta(\tau)$. The parameters of the impulse response, $a$ and $T_R$, are a constant amplitude term and the reverberation time, respectively [3].

In this model, the power envelope of the reverberant signal, $e_y(t)^2$, can be determined as

$$\langle y(t)^2 \rangle = e_x(t)^2 + e_h(t)^2 = e_y(t)^2,$$  \hspace{1cm} (4)

where $\langle \cdot \rangle$ is a set-averaging operation [3]. This equation shows that a significant relationship exists between the envelopes; i.e. the MTF concept. Based on this result, $e_x(t)^2$ can be recovered by deconvoluting $e_y(t)^2$ with $e_h(t)^2$. Here, the transmission functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to be the $z$-transforms of $e_x(t)^2$, $e_h(t)^2$, and $e_y(t)^2$, respectively. Thus, the transmission function of the power envelope of the original signal, $E_x(z)$, can be determined from

$$E_x(z) = \frac{E_y(z)}{a^2} \left\{ 1 - \exp \left( -\frac{13.8}{T_R} z^{-1} \right) \right\},$$  \hspace{1cm} (5)

where $f_s$ is a sampling frequency. Finally, the power envelope $e_x(t)^2$ can be obtained from the inverse $z$-transform of $E_x(z)$ [3]. This model concept, therefore, enables recovery of the power envelope of an original signal from an observed signal.

However, two problems are associated with the basic model: (1) how to precisely extract the power envelope from the observed signal, and (2) how to determine the parameters of the reverberation time and the amplitude terms ($T_R$ and $a$) of the impulse response of the room acoustics. We have resolved these problems, as previously reported [5].

(a) Extraction of the power envelope:

$$\hat{e}_y(t)^2 = \text{LPF} \left[ < y(t) \hat{n}(t)^2 > \right].$$  \hspace{1cm} (6)

(b) Estimation of the reverberation time:

$$T_R = \max \left\{ \arg\min_{0 \leq T_R \leq T_{R,\text{max}}} \left\{ \int_0^T |{\hat{e}_x}_{T_R}(t)^2, 0| \, dt \right\} \right\}.$$  \hspace{1cm} (7)

(c) Determination of the amplitude term:

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

We used LPF(\%) of the low-pass filtering with a cut-off frequency of 20 Hz to remove the high-pass envelope [5]. Here, $T$ is a signal duration of $y(t)$, $\hat{e}_x(T_R)^2(t)$ is the set of candidates of the power envelope dereverberated as a function of $T_R$, and $T_{R,\text{max}}$ is the upper limited region of $T_R$.

![Figure 2: Power envelope inverse filtering method in the constant bandwidth filterbank model.](image)

![Figure 3: An example of power envelopes based on the MTF and improved estimation of $T_R$: (a) Envelopes with no silence, (b) Envelope with a long silence, (c) Candidates of the dereverberated envelopes, and (d) Threshold points, $T_R(t)$.](image)

In this paper, we extend the improved model [5] into the filterbank model for speech applications because most of the envelopes are not co-modulated envelopes throughout the entire range. This model is shown in Fig. 2. Here, a constant bandwidth filterbank is used to decompose speech into channel components that are then dereverberated using the same method in each channel.

3.2. Application problems

There are still two problems that apply this method to speech.

The first problem is determining the appropriate bandwidth. In the model concept, there is a trade-off regarding bandwidths in the filterbank. Bandwidths have to be narrow to enable a precise analysis of the speech envelope, but they have to be wide to satisfy Eq. (4) in the MTF concept. We therefore have to determine a reasonable channel bandwidth. We examined the correlation between the power envelopes on channels in a narrowband (40 Hz) filterbank to verify the co-modulation characteristics, and found that an appropriate bandwidth was less than or equal to 500 Hz while the correlation was over 0.9. On the other hand, we also examined the correlation between $e_y(t)^2$ obtained from Eq. (4) and $\hat{e}_y(t)^2$ extracted from $y(t)$ using Eq. (6). In this case, we found the appropriate bandwidth was greater than about 300 Hz while the correlation was over 0.9. Based on the trade-off, in this paper we consider 400 Hz to be a reasonable bandwidth in the filterbank.

The second problem is the need for a better method to estimate $T_R$ in when there is a long silence in the envelope. For example, Fig. 3 (a) shows a sinusoidal envelope of 10 Hz (solid line) and a reverberator envelope with $T_R = 0.5$ (dashed line).
In this figure, we can precisely estimate $\hat{T}_R$ using Eq. (7) and then dereverberate $\hat{c}_e(t)^2$ from $\hat{v}_e(t)^2$ using Eq. (5). However, this processing cannot work in the case shown in Fig. 3 (b). This is because we assumed that the modulation index of the original signal was set to 1 and a silence interval corresponding to the modulation index of 1 was not long [5], and therefore the processing will not work if this assumption is not satisfied. In this paper, instead of focusing on the negative area in Eq. (7), we focus on the variations of shifting at a related position on the same threshold, depending upon inverse filtering with $T_R$, as shown in Fig. 3 (c) (denoted by ‘~’). The dotted lines show the over- and/or under-dereverberated envelopes when using various $T_R$. Thus, we propose an improved method for estimating $T_R$ even if there is a long silence, as follows.

$$T_P(T_R) = \min \left( \arg \min_{T_{\min} \leq T \leq T_{\max}} \left| \hat{v}_e(T, T_R) - \theta \right| \right),$$  \hspace{1cm} (9)

$$\hat{T}_R = \arg \min_{0 \leq T_R \leq T_{\max}} \left\{ \frac{dT_P(T_R)}{dT_R} \right\},$$ \hspace{1cm} (10)

where $\theta$ is a threshold for detecting a point from the maximum of $v_e(t)^2$. Here, we set $\theta$ to −20 dB from the maximum envelope. $T_{\min}$ and $T_{\max}$ are, respectively, the lower and the upper limited regions for determining a point.

Figure 3 (d) shows the variation of the shifting point with any $T_R$. In this figure, $T_P$ at $T_R = 0.2$ could identify the correct $T_R = 0.5$ where we detected a rapid variation at $T_P(T_R)$. Therefore, we can precisely estimate $T_R$ using Eq. (10).

4. Carrier regeneration

Figure 4 shows the signal flow of the carrier regeneration block. This block separately processes in voiced and unvoiced durations based on the estimated F0. In this paper, we assume that F0 has been estimated accurately in this block.

First, the carrier regeneration block in the voiced duration is based on the facts: source information of voiced speech consists of harmonicity and/or periodicity with the fundamental frequency and this harmonicity can smoothly vary with the F0 fluctuation; the harmonicity is frequency-modulated with F0 together. Thus, this block generates harmony with the estimated F0 based on the PIFM model [6]. In this block, the generated carrier $\hat{c}_e(t)$ is represented as

$$\hat{c}_e(t) = \frac{1}{K(t)} \sum_{k=1}^{K(t)} \sin \left( k \int_0^t 2\pi F_0(\tau) d\tau + \phi_k(t) \right),$$ \hspace{1cm} (11)

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 block in the unvoiced duration generates a white noise carrier $\hat{c}_n(t)$, as shown in Fig. 4. All carriers corresponding to a speech signal are then added together, i.e., $c_e(t) = c_e(t) + c_n(t)$. Finally, this carrier is decomposed into carriers in each channel of the same filterbank in order to multiply carriers with the dereverberated envelopes. Carriers are then reconstructed into dereverberated speech using the same filterbank as is used for speech-resynthesis.

5. Simulations

We carried out simulation as follows to evaluate the proposed model. The speech signals were three Japanese sentences ('laikawazuzu', 'shinbin', and 'joudan') uttered by ten speakers (five males: Mau, Mht, Mnm, Mtm, and Mtt and five females: Faf, Fis, Fkn, Fsu, and Fyn) from the ATR-database [7], 100 types of impulse response $h(t)$ were used, and five reverberation times $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ were used. All stimuli, $y(t)$s, were composed through $15,000 (=3 \times 10 \times 5 \times 100)$ convolutions of $x(t)$ with $h(t)$s.

First, we evaluated the power envelope dereverberation in each channel. Figure 5 shows the improved correlation and the improved SNR of each channel for the dereverberation from the speech signals, comparing $\varepsilon_e(t)^2$ with $\varepsilon_e(t)^2$ and $\varepsilon_e(t)^2$ with $\varepsilon_e(t)^2$, in this figure, the bar height and the error bar show the mean and the standard deviation of each result, respectively. We found that the improved correlation and the improved SNR can be increased with increasing $T_R$, except with $T_R$ of 0.1.
We proposed a speech dereverberation method, based on the MFT concept, which does not require measurement of the impulse response of room acoustics. In this model, the power envelopes and carriers are decomposed from the reverberant speech signal using an $N$-channel filterbank and then are dereverberated in each respective channel. In the envelope dereverberation, we use a power envelope inverse filtering method based on the MFT concept. In the carrier regeneration, the carrier generation method based on voiced/unvoiced intervals from the estimated fundamental frequency based on the PIFM method. In this paper, we assumed that the F0 is known. Through 15,000 simulations of the dereverberation of reverberant speech, we found that the proposed model can accurately dereverberate both reverberant speech and the power envelopes.

In our future work, we will (1) reconsider an adaptive dereverberation method based on the time-division and frequency-division processing using the reconstructed filterbank depending on each speech signal, and (2) attempt to construct a method to estimate F0 from reverberant speech.

### 5. Conclusions

We proposed a speech dereverberation method, based on the MFT concept, which does not require measurement of the impulse response of room acoustics. In this model, the power envelopes and carriers are decomposed from the reverberant speech signal using an $N$-channel filterbank and then are dereverberated in each respective channel. In the envelope dereverberation, we use a power envelope inverse filtering method based on the MFT concept. In the carrier regeneration, the carrier generation method based on voiced/unvoiced intervals from the estimated fundamental frequency based on the PIFM method. In this paper, we assumed that the F0 is known. Through 15,000 simulations of the dereverberation of reverberant speech, we found that the proposed model can accurately dereverberate both reverberant speech and the power envelopes.

In our future work, we will (1) reconsider an adaptive dereverberation method based on the time-division and frequency-division processing using the reconstructed filterbank depending on each speech signal, and (2) attempt to construct a method to estimate F0 from reverberant speech.

### 6. Acknowledgment

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

### 8. References