A METHOD BASED ON THE MTF FOR RECOVERING THE POWER ENVELOPE FROM REVERBERANT SPEECH

PACS REFERENCE: 43.72.Ew

UNOKI Masashi, FURUKAWA Masakazu, and AKAGI Masato
School of Information Science, Japan Advanced Institute of Science and Technology
Asahidai, Tatsunokuchi, Nomi, Ishikawa 923-1292 JAPAN
Tel: +81-761-51-1237
Fax: +81-761-51-1149
E-mail: {unoki, m-furuka, akagi}@jaist.ac.jp

ABSTRACT This paper proposes a method for recovering the power envelope from reverberant speech. This method is based on the modulation transfer function (MTF) and does not require that the impulse response of an environment be measured. It improves upon the basic model proposed by Hirobayashi et al. regarding the following problems: (i) how to extract the power envelope from the observed signal; (ii) how to determine the parameters of the impulse response of the room; and (iii) application of the MTF to speech and the anti-co-modulation of the speech envelope. We have shown that the proposed method can accurately dereverberate the power envelope from reverberant speech.

1. INTRODUCTION

Recovery of the original signal from a reverberant signal is an important issue concerning speech signal processing such as preprocessing for speech recognition.

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 constraining non-overlaps of zeros in all pairs of the impulse responses between the sources and the microphones. This method can be applied to room acoustics with non-minimum phase characteristics. However, these methods have to measure the impulse response of the room 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.

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. However, three problems remain concerning this method: how to extract the power envelope, how to determine the model parameters, and whether it can be applied to speech.

In this paper, we propose a method, also based on the MTF, for recovering the power envelope from reverberant speech that resolves these issues.

2. THE POWER ENVELOPE INVERSE FILTERING METHOD

2.1. Model concept based on the MTF

In the Hirobayashi et al.'s model, the original signal and the stochastic-idealized impulse response in the room [4] are assumed to be $x(t)$ and $h(t)$, respectively, and these are modeled based on the MTF as follows [3]:

$$x(t) = e_x(t) n_1(t), \qquad (1)$$

$$h(t) = e_h(t) n_2(t) = a \exp\left(-6.9t/T_R\right) n_2(t), \qquad (2)$$

$$\left\langle n_k(t), n_k(t-\boldsymbol{t})\right\rangle = \boldsymbol{d}(\boldsymbol{t}), \qquad (3)$$

where $e_x(t)$ and $e_h(t)$ are the power envelopes of $x(t)$ and $h(t)$, and $n_1(t)$ and $n_2(t)$ are the mutually independent respective white noise values. Parameters of the impulse response, a and T_R, are a constant amplitude term and the reverberation time, respectively [3].

In this model, the reverberant signal $y(t)$ is the convolution of $x(t)$ with $h(t)$ in the time domain, so the power envelope of $y(t)$ can be determined as

$$\left\langle y(t)^2 \right\rangle = \left\langle \left\{ \int_{-\infty}^{\infty} x(\boldsymbol{t}) h(t-\boldsymbol{t}) d\boldsymbol{t} \right\}^2 \right\rangle,$$

$$= \int_{-\infty}^{\infty} e_x(t)^2 e_h(t-\boldsymbol{t})^2 d\boldsymbol{t} = e_y(t)^2. \qquad (4)$$

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 $E_y(z)$,

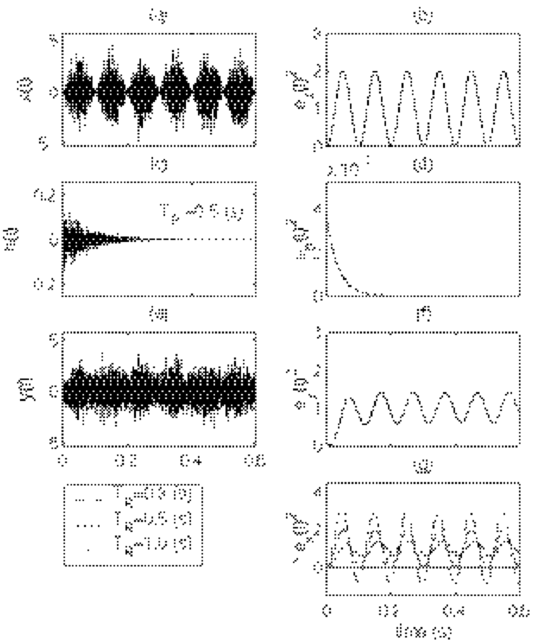

Figure 1. Example of the relationship between the power envelopes of a system based on the MTF concept.

are assumed to be the z-transforms of $e_h(t)^2$, and $e_y(t)^2$, respectively. Thus, the transmission function of the power envelope of the original signal can be determined from

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

where f_s is the sampling frequency. Finally, the power envelope $e_x(t)^2$ can be determined from the inverse z-transform of $E_x(z)$ [3].

Figure 1 shows an example of how the power inverse filtering method is related to the MTF concept. Figure 1(a) shows the original signal with a sinusoidal power envelope (Fig. 1(b), the modulation frequency was 10 Hz and the modulation index was 1). Figure 1(c) shows the impulse response of Eq. (2) with $T_R = 0.5$, and Fig. 1(e) shows the observed signal with a convolution of $x(t)$ with $h(t)$. The right panels ((b), (d), and (f)) show the power envelopes of the signals. The MTF concept is to show the modulation index as a function of the modulation frequency and the reverberation time [4].

The solid line in Fig. 1(g) shows the power envelope recovered from the observed power envelope (Fig. 1(f)) through this method with $T_R = 0.5$. If the method is applied with T_R set to an inappropriate value, the recovered power envelopes are not precisely dereverberated as the other lines in Fig. 1(g) show.

2.2. Problems

With this concept, the basic model can dereverberate the power envelope of an original signal from an observed signal if it can detect the power envelope precisely and the parameters of the room impulse response are known before processing, as shown in Fig. 1(g). However, we still have to overcome the problems associated with the basic model. Here, we consider the first two: (1) how to precisely extract the power envelope from the observed signal, and (2) how to determine the parameters of the reverberant time and the amplitude terms (a and T_R) of the impulse response.

3. IMPROVED MODEL

3.1. Extraction of the power envelope

Extracting the power envelope from an observed signal based on the MTF concept using well-known techniques (such as the half-wave rectification (HWR) of the signal demodulation) is difficult because the carrier is white noise rather than a sinusoidal signal.

In this paper, we propose two methods that can be used to extract the power envelope. One is a method using set-averaging:

$$\hat{e}_y(t)^2 = \left\langle \hat{\mathbf{y}}(t)^2 \right\rangle = \mathrm{LPF}\left[\left\langle \left(y(t)\mathbf{n}(t) \right)^2 \right\rangle\right]. \qquad (6)$$

The other is a method using the Hilbert transform:

$$\hat{e}_y(t)^2 = \mathrm{LPF}\left[\mathrm{Hilbert}\left(y(t)\right)^2\right]. \qquad (7)$$

In both equations, we used low-pass filtering (LPF) as post-processing to remove the high-pass envelope. In this paper, we use an LPF cut-off frequency of 20 Hz because an important modulation region for speech perception and speech recognition is from 1 to 16 Hz [6].

3.2. Determination of the impulse response parameters

In Hirobayashi et al.'s model, they did not describe how to determine parameter of a. In this model, however, we find that a is given the same value for both Eqs. (2) and (5), so that it may be no critical problem. Since a is related to the gain of the room acoustics, we assume that the value of a determined from the summarized $e_h(t)^2$ is 1 for applications in real environment.

On the one hand, parameter T_R must be precisely determined from the observed signal for dereverberation. Hirobayashi et al. used the known T_R in their model [3], so that model is restricted for any application.

In this paper, we consider over- and/or under-recovery of the power envelope with T_R as shown in Fig. 1(g). A matching-condition of the original and recovered power envelope is to recover a modulation index of 1 from the reverberation if the modulation index of the original signal is assumed to be 1. This condition can be satisfied by detecting a timing-point where the minimum dip will be 0 or the negative area of the recovered envelope will be 0. In this paper, we assume that the modulation index of the original signal is set to 1, so T_R can be estimated using

$$\hat{T}_R = \underset{T_{R,\min} \le T_R \le T_{R,\max}}{\mathrm{argmax}} \left\{ \int_0^T \min\left(\hat{e}_{x,T_R}(t)^2,\, 0 \right) dt \right\}, \qquad (8)$$

where $\hat{e}_{x,T_R}(t)^2$ is defined as a function of T_R. Here, $T_{R,\min}$ is the lower limited region of T_R and is determined from a timing-point where a negative area is arisen from. $T_{R,\max}$ is the upper limited region. If the original signal has more silences in signal duration, this assumption is reasonable because more zero-dips exist in the power envelope.

3.3. Evaluation

In this section, we evaluate the improved model. Figure 2 shows a block diagram of the power envelope inverse filtering. The values of $x(t)$ consisted of the white noise multiplied by the three types of power envelope:

(a) Sinusoidal: $e_x(t)^2 = 1 - \cos(2\pi Ft)$,

(b) Harmonics: $e_x(t)^2 = 1 + \frac{1}{K}\sum \sin(2\pi kt + \theta_k)$,

(c) Band-limited noise: $e_x(t)^2 = \mathrm{LPF}[n(t)]$.

Here, $F = 15$ Hz, f_s = 20 kHz, $K = 20$, and θ_k is a random phase. The impulse responses, $h(t)$, consisted of five types of envelope, with $T_R = 0.1,\ 0.3,\ 0.5,\ 1.0,\ 2.0$, multiplied by 100 white noise carriers. All stimuli, $y(t)$, were composed through 1,500 (= $3\times 5\times 100$) convolutions of $x(t)$ with $h(t)$.

As evaluation measures, we used (1) the correlation and (2) the SNR (where S is the original signal and N is the difference between S and the estimated signal) between the original envelope and the extracted/recovered envelope.

Figure 3 shows the extraction accuracy for the power envelopes from all stimuli using the set-averaging method (at point A in Fig. 2). Each point and the error bar show the mean and the standard deviation of the results. We found that the proposed method can precisely extract the power envelope from the observed signal. The method using the Hilbert transform could also extract it as accurately as the set-averaging method, but the HWR method could not.

Figure 4 shows T_R (at point B in Fig. 2) estimated using the results of Fig. 3. Each point and the error bar show the mean and the standard deviation for $\hat{T}_R$. The dotted line shows the idealized $\hat{T}_R$. We found that $\hat{T}_R$ matched the idealized value from 0 to about 0.5, but there were discrepancies with the idealized value above about 0.5.

Figure 5 shows the improvement in accuracy for the dereverberation (at point C in Fig. 2), obtained by plotting the differences between the recovered power envelope with and without the model. The improvements in Fig. 5 are positive values, demonstrating that the proposed model could effectively dereverberate the power envelope of the signal from the observed signal. We

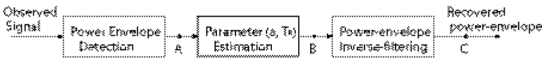

Figure 2. Block diagram of the power envelope inverse filtering method.

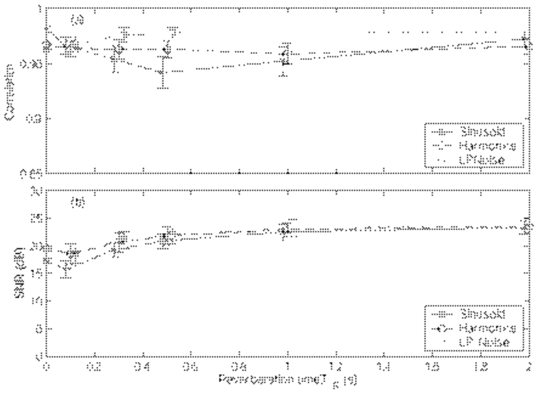

Figure 3. Extraction accuracy of the power envelope: (a) correlation, and (b) SNR.

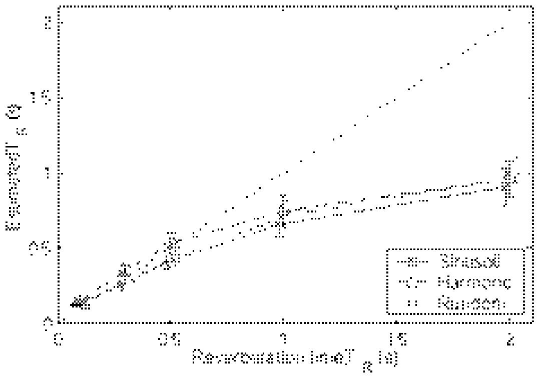

Figure 4. Estimated reverberation time. The dotted line shows the idealized reverberation time.

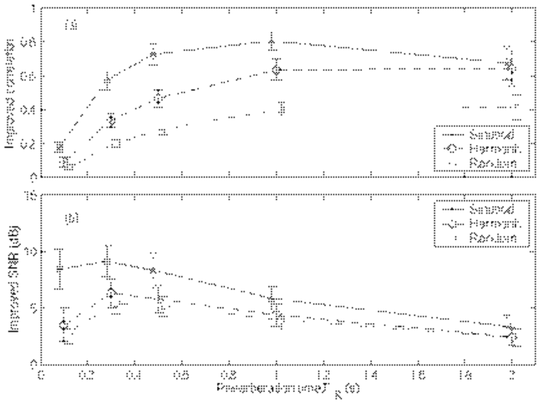

Figure 5. Improvement of dereverberation accuracy in the proposed model: (a) correlation, and (b) SNR.

compared these results with the result when $\hat{T}_R$ was set to a known value (the idealized value). In this case, there were no differences in the improvements with the proposed model when T_R was from 0 to about 0.5, although the improvements in the SNR fell by about 1 dB at $T_R = 1.0$ and by about 3 dB at $T_R = 2.0$. Therefore, Eq. (8) can be taken as a reasonable constraint for dereverberation of the power envelope in this model.

4. APPLICATION TO SPEECH

4.1. Application Considerations

In the previous section, we showed how the improved model solves some problems concerning the basic model proposed by Hirobayashi et al. In this section, we consider the model's application to reverberant speech as regards three points.

The first is whether the MTF concept can be applied to speech. To apply this concept, we should ensure that carriers are not correlated with each other, but speech carriers may not remain uncorrelated. Let us consider the modeling in terms of this difference. Figure 6 shows a result of dereverberation using the proposed model for the same stimuli, except with carriers, as in Fig. 5. The carriers were 100 types of harmonics with F0 of 100 Hz and random phases. We found that the proposed model could recover the power envelope from the observed signal in this case as well as in Fig. 5, although there was a large deviation.

The second point to consider is whether the power envelopes of speech for all frequencies have a co-modulation characteristic. If they do, the third point is what is an appropriate bandwidth that can be regarded as similar to the co-modulation for speech.

We examined the correlation between the power envelopes on channels in a filterbank to verify the co-modulation. Figure 7 shows an example of this analysis for speech using a constant narrow-band (40 Hz) filterbank. The speech signal was a Japanese sentence (/aikawarazu/) uttered by a male speaker in the ATR-database [7]. In this figure, the contour shows the region of correlation over 0.95 as co-modulation. Our results show that the 400-Hz-bandwidth power envelopes can be interpreted as co-modulation although the power envelope over all regions cannot be interpreted in this way. We also found that there is a reasonable trade-off between the divided co-modulation bandwidths and the minimum bandwidth to be held for the MTF. Note that the MTF cannot be held on a band that is too narrow because the assumption of noncorrelation is not held consistently.

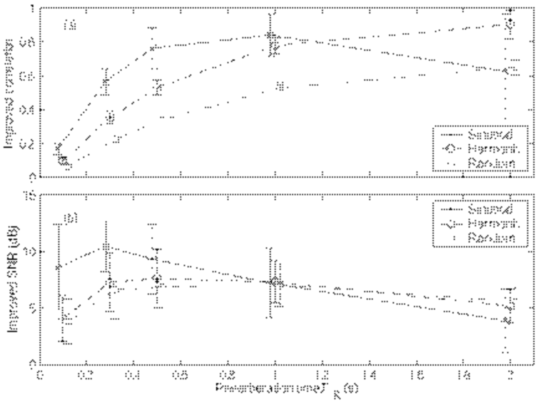

Figure 6. Improvement of the dereverberation accuracy: (a) correlation, and (b) SNR. (Carriers are harmonics.)

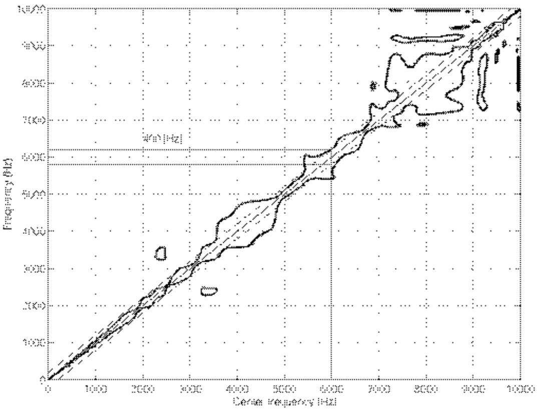

Figure 7. Correlation between the power envelopes of one channel and the adjacent channels for a speech n filterbank.

4.2. Extended method on filterbank

Based on the above considerations, we extended the proposed model into the filterbank model for speech. Figure 8 shows the architecture of the extended model. In this model, the bandwidth of each channel is set to 400 Hz and the extraction method using the Hilbert transform is employed for computation cost. The blocks where T_R is estimated are processed separately.

Figure 9 shows the improved correlation and the improved SNR of each channel (the bar height represents the mean of each result) for the dereverberation from the same speech signal as in Fig. 7 (100 types of impulse response and five reverberation times). These results demonstrate that the proposed model can be used to precisely recover the power envelopes from reverberant speech.

5. CONCLUSION

We have developed a method for recovering the power envelope from reverberant speech based on the MTF without measuring the impulse response of an environment. This method improves upon an earlier method in several ways: (i) extraction of the power envelope from the observed signal; (ii) determination of the parameters (T_R and a) of the impulse response; (iii) applicability of the MTF concept to speech; and (iv) extension of the concept to the filterbank model by taking the modulation in consideration. We have shown that our method can be used to accurately dereverberate the power envelope from reverberant speech.

ACKNOWLEDGEMENTS

This work was supported by CREST of JST and by a grant-in-aid for scientific research from the Ministry of Education (No. 14780267).

REFERENCES

[1] Neely, S. T. and Allen, J. B. “Invertibility of a room impulse response,” J. Acoust. Soc. Am. Vol. 66, No. 1, July 1979.

[2] Miyoshi, M. and Kaneda, Y., “Inverse filtering of room acoustics,” IEEE Trans. ASSP, Vol. 36, No. 2, pp. 145-152, Feb. 1988.

[3] Hirobayashi, S. et al. “Speech waveform recovery from a reverberant speech signal using inverse filtering of the power envelope transfer function,” Trans. IEICE A, Vol. J81-A, No.10, pp. 1323-1330, Oct. 1998. (in Japanese)

[4] Houtgast, T., Steenken, H.J.M., and Plomp, R., “Predicting speech intelligibility in room acoustics,” Acoustica, Vol. 46, pp. 60-72, 1980.

[5] Kanedera, N. et al., “On the importance of various modulation frequencies for speech recognition,” Proc. EuroSpeech97, pp. 1079-1082, Rhodes, Greece, Sept. 1997.

[6] Takeda, K. et al., Speech Database User’s Manual, ATR Technical Report TR-I-0028, 1988.

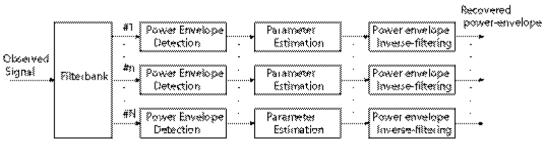

Figure 8. Power envelope inverse filtering in the constant bandwidth filterbank.

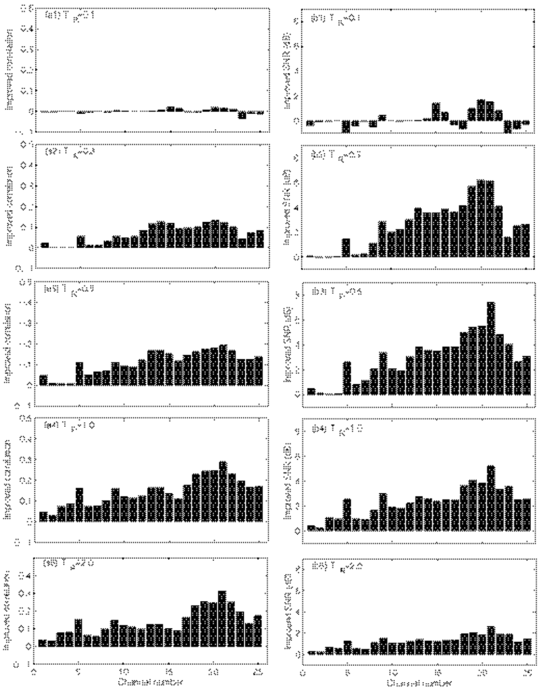

Figure 9. Improvement of dereverberation accuracy for the power envelope of speech on the filterbank.