MTF-BASED POWER ENVELOPE RESTORATION IN NOISY REVERBERANT ENVIRONMENTS

Masashi Unoki, Yutaka Yamasaki, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 JAPAN
email: {unoki, yutaka1017, akagi}@jaist.ac.jp

ABSTRACT

Many speech enhancement methods had been proposed to suppress the effects of noise or reverberation. Although most of methods aimed to enhance only noisy or reverberant speech, these cannot simultaneously enhance noisy reverberant speech. This paper proposes a method for restoring the power envelope from the noisy reverberant speech. This method is based on the MTF concept and does not require that the impulse response and noise conditions in the room acoustics (noisy reverberant environments) be measured. The proposed method suppresses the effects of reverberation and noise on the power envelopes by restoring the smeared MTF. We carried out 12,000 simulations of noise-suppression and dereverberation for noisy reverberant speech to objectively evaluate the proposed method. The results showed that the proposed method can simultaneously work well in both noise-suppression and dereverberation.

1. INTRODUCTION

In real environments, significant features of speech are smeared because of noise and reverberation. The quality of sound and intelligibility of speech are significantly reduced. We need to improve noisy reverberant speech (noise suppression and dereverberation) in various speech signal processing systems, such as hearing aids and the preprocessing for automatic speech recognition (ASR).

There are several well-known suppression methods that can be used to remove the effects of noise or reverberation in either noisy or reverberant environments. For example, the spectral subtraction method proposed by Boll [1], the Kalman filtering method proposed by Paliwal and Basu [2], minimum-phase inverse filtering method proposed by Neely and Allen [3], and the multiple input/output inverse theorem (MINT) method proposed by Miyoshi and Kaneda [4]. Although these methods can work well in either noisy (first two methods) or reverberant environments (last two methods), they cannot simultaneously work well these environments.

Recently, Kinoshita *et al.* have found a way to enhance speech recorded in noisy reverberant environments, by considering two sequential processes: noise reduction using spectral subtraction for noisy reverberant speech and then dereverberation using linear prediction for noise-reduced reverberant speech [5]. Although this seems to be the simplest modeling of the effects of noise and reverberation, it is thought that a combination of different systems (representations and/or processing) cannot simultaneously deal with the effects of additive noise and reverberation.

On the other hand, Houtgast and Steeneken have proposed a method of prediction that can assess the effects of the enclosure on the intelligibility of speech in both noisy and reverberant environments by using the modulation transfer function (MTF) [6]. The MTF concept enables noise and reverberation to be simultaneously suppressed.

We had already proposed the use of a power envelope inverse filtering method that is based on the MTF concept [7, 8]. Using our method restored the power envelope of the reverberant speech, recovered 30% of the loss of intelligibility caused by the reverberation [9] and produced 30% relative improvement in the error reduction rate in ASR due to reverberation [10], under the only noise-less condition. If we could extend this method as a noise suppression method based on the MTF concept, we would be able to propose the use of a way of using MTF to enhance speech, which suppresses noise and reverberation simultaneously. The goal of our work is to propose a method of enhancing speech that can be used to reduce the effects of noise and reverberation. As our first step, we proposed restoring the smeared MTF to restore the power envelopes of noisy reverberant speech.

2. MTF CONCEPT

The MTF concept was proposed by Houtgast and Steeneken to account for the relation between the degree of modulation of the envelopes of input and output signals and the characteristics of the enclosure and a way to predict the speech transfer index, which is strongly related to intelligibility [6]. This concept was introduced as a measure in room acoustics for assessing the effect of the enclosure on intelligibility. They defined input and output temporal power envelopes as

$$\mathbf{Input} = \overline{I_i^2}(1+\cos(2\pi f_m t)), \quad (1)$$

$$\mathbf{Output} = \overline{I_o^2}\{1+m(f_m)\cos(2\pi f_m(t-\tau))\}, \quad (2)$$

where $\overline{I_i^2}$ and $\overline{I_o^2}$ are the input and output intensities, f_m is the modulation frequency, and τ is the phase information. The modulation index of the power envelope is $m(f_m)$ and referred to as MTF. We will now explain the MTF in noisy and/or reverberant environments.

2.1 Model concept based on the MTF

We assume the output, the input, the impulse response, and the noise signals to be $y(t)$, $x(t)$, $h(t)$, and $n(t)$. These are modeled based on the MTF [7, 8, 9] as follows:

$$y(t) = h(t)*x(t)+n(t), \quad (3)$$

$$x(t) = e_x(t)c_x(t), \quad (4)$$

$$h(t) = e_h(t)c_h(t) = a\exp(-6.9t/T_R)c_h(t), \quad (5)$$

$$n(t) = e_n(t)c_n(t), \quad (6)$$

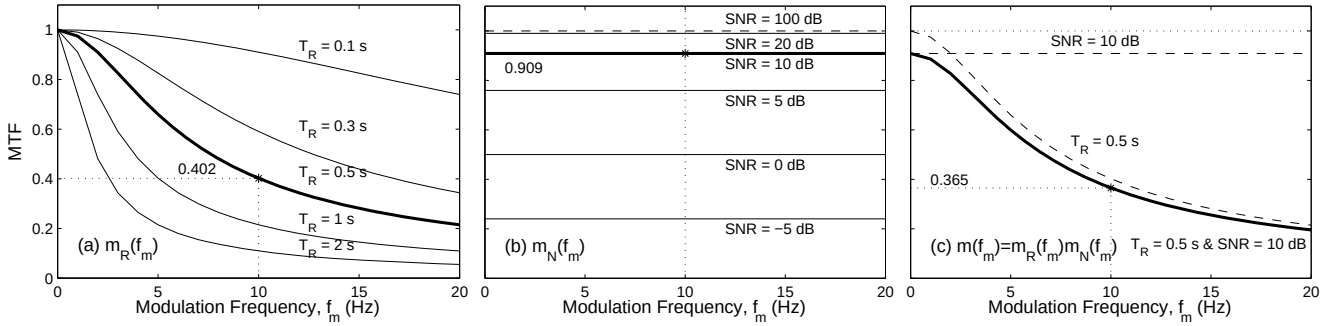

Figure 1: Theoretical representations of the MTFs, $m(f_m)$, in (a) reverberant environment, (b) noisy environment, and (c) both noisy and reverberant environments. The bold solid lines indicate the MTF with $T_R = 0.5$ s and SNR $= 10$ dB.

where $e_x(t)$, $e_h(t)$, and $e_n(t)$ are the temporal envelopes of $x(t)$, $h(t)$, and $n(t)$. $c_x(t)$, $c_h(t)$, and $c_n(t)$ are carriers such as random variable. Here, $\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau)$ and $\langle\cdot\rangle$ is an ensemble average operation. T_R is the reverberation time. In this model, $e_y^2(t)$ can be derived as

$$\langle y^2(t)\rangle = \langle h^2(t) * x^2(t)\rangle + \langle n^2(t)\rangle, \quad (7)$$

$$e_y^2(t) = e_h^2(t) * e_x^2(t) + e_n^2(t). \quad (8)$$

(see [7, 8] for a detailed derivation of Eq. (8)). We used the relationship between temporal power envelopes to restore $e_x^2(t)$ from the observed $e_y^2(t)$.

2.2 MTF in reverberant environments

In the reverberant condition, the input and output temporal power envelopes, $e_x^2(t)$ and $e_y^2(t)$, are represented as

$$e_x^2(t) = \overline{e_x^2}(1+\cos(2\pi f_m t)), \quad (9)$$

$$e_y^2(t) = e_x^2(t) * e_h^2(t) = \frac{\overline{e_x^2}}{\alpha}\{1 + m_R(f_m)\cos(2\pi f_m t)\}, \quad (10)$$

where $\alpha = \int_0^\infty h^2(t)dt$ and $\beta = \int_0^\infty h^2(t)\exp(-j\omega_m t)dt$. The complex MTF in reverberant environments is defined as

$$m_R(f_m) = \left|\frac{\beta}{\alpha}\right| = \sqrt{1+\left(2\pi f_m \frac{T_R}{13.8}\right)^2} \quad (11)$$

The MTF in reverberant environments depends on f_m. This means the low-pass characteristics as a function of T_R (as shown in Fig. 1(a)). In the case of a T_R of 0.5 s, $m(f_m)$ at $f_m = 10$ Hz is 0.402.

2.3 MTF in noisy environments

Where there is additive noise, $e_y^2(t)$ is represented as

$$e_y^2(t) = e_x^2(t) + e_n^2(t) = \left(\overline{e_x^2} + \overline{e_n^2}\right)\{1 + m_N(f_m)\cos(2\pi f_m t)\}, \quad (12)$$

where $\overline{e_n^2} = \frac{1}{T}\int_0^T e_n^2(t)dt$. Here, $e_n^2(t)$ is assumed to be constant in the time domain and T is the signal duration. The complex MTF in noisy environments, is defined as

$$m_N(f_m) = \frac{\overline{e_x^2}}{\overline{e_x^2} + \overline{e_n^2}} = \frac{1}{1+10^{-(\mathrm{SNR})/10}}, \quad (13)$$

where $\mathrm{SNR} = 10\log_{10}(\overline{e_x^2}/\overline{e_n^2})$ in dB. This MTF is independent of f_m and reduced as a function of SNR (Fig. 1(b)). In the case of SNR of 10 dB, $m(f_m)$ is 0.909.

2.4 MTF in noisy and reverberant environments

The MTF in noisy reverberation environments calculated from Eqs. (11) and (13), can be represented as

$$m(f_m) = m_R(f_m)\cdot m_N(f_m) = \sqrt{1+\left(2\pi f_m \frac{T_R}{13.8}\right)^2} \Big/ \left(1+10^{-\frac{\mathrm{SNR}}{10}}\right). \quad (14)$$

The MTF in noisy reverberant environments depends on f_m. This means the low-pass characteristics resulting from reverberation as a function of T_R and the constant attenuation resulting from noise as a function of SNR (Fig. 1(c)). In the case of a T_R of 0.5 s and SNR $= 10$ dB, $m(f_m)$ at $f_m = 10$ Hz is 0.365 ($= 0.402 \times 0.909$). Hence, the effect of noise and reverberation can be suppressed by using the inverse filtering of MTF in Eq. (14).

3. POWER ENVELOPE RESTORATION

3.1 Implementation

We propose the use of the power envelope restoration based the MTF concept. A block-diagram of the method is shown in Fig. 2. This method consists of (i) power envelope extraction, (ii) power envelope subtraction, and (iii) power envelope inverse filtering with parameter estimation. Here, the constant bandwidth filterbank was used to analyze the signal.

The power envelopes from $y(t)$ are extracted by

$$e_y^2(t) = \mathrm{LPF}\left[\left|y(t) + j\mathrm{Hilbert}[y(t)]\right|^2\right], \quad (15)$$

where LPF$[\cdot]$ is a low-pass filtering and Hilbert$[\cdot]$ is the Hilbert transform. This method is based on a calculation of the instantaneous amplitude of the signal, and is used in low-pass filtering as post-processing to remove the higher frequency components in the power envelopes. We used LPF with a cut-off frequency of 20 Hz [7, 8, 9].

Power envelope subtraction on the basis of the MTF concept is done to suppress the noise. The modulation index and the averaged power in Eq. (13) are only affected by noise. To restore the first term in Eq. (8) from the power envelope of

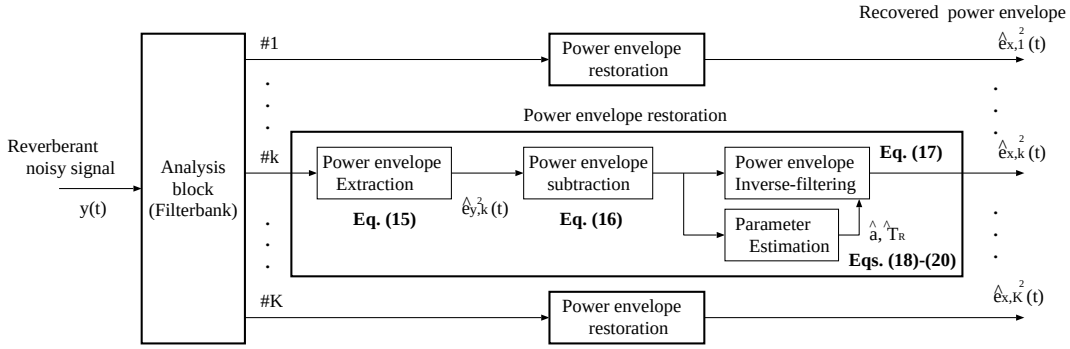

Figure 2: The power envelope restoration method.

noisy reverberant signal $e_y^2(t)$, $m_N(f_m)$ is utilized as follows.

$$\begin{aligned}\hat{e}_x^2(t) &= \overline{e_x^2}\left(1+m_N(f_m)\cos(2\pi f_m t)\times\frac{1}{m_N(f_m)}\right),\\ &= e_y^2(t)-\overline{e_n^2}. \end{aligned} \quad (16)$$

Here, the robust VAD method (e.g., [12]) is used to calculate $\overline{e_n^2}$ (N) and $(\overline{e_x^2}+\overline{e_n^2})$ (SN) in Eq. (13) from the observed $e_y^2(t)$ in noise duration and signal+noise duration respectively.

On the basis of this result, $e_x^2(t)$ can be recovered by deconvoluting $e_y^2(t)=e_x^2(t)*e_h^2(t)$ in Eq. (8) with $e_h^2(t)$. 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 $e_x^2(t)$, $e_h^2(t)$, and $e_y^2(t)$. Thus, the $E_x(z)$ can be determined from

$$E_x(z)=\frac{E_y(z)}{a^2}\left\{1-\exp\left(-\frac{13.8}{T_R\cdot f_s}\right)z^{-1}\right\}, \quad (17)$$

where f_s is the sampling frequency. The power envelope $e_x^2(t)$ can then be obtained from the inverse z-transform of $E_x(z)$. Here, two parameters (T_R and a) are obtained as [9]

$$\hat{T}_R = \underset{0\le T_R\le T_{R,\max}}{\arg\min}\left\{\frac{dT_P(T_R)}{dT_R}\right\}, \quad (18)$$

$$T_P(T_R) = \min\left(\underset{t_{\min}\le t\le t_{\max}}{\arg\min}\left|\hat{e}_{x,n,T_R}(t)^2-\theta\right|\right), \quad (19)$$

$$\hat{a} = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt}. \quad (20)$$

3.2 Example

An example of how the power envelope restoration is related to the MTF concept is shown in Fig. 3. A sinusoidal power envelope as the original $e_x^2(t)$ ($=0.5(1+\sin(2\pi f_m t))$) and $x(t)$ calculated from $e_x^2(t)$ and a white noise carrier $c_x(t)$ using Eq. (4) are shown in Figs. 3(a) and (b); f_m was 10 Hz and $m(f_m)$ was 1. Figures 3(c) and (d) show $e_h^2(t)$ with $T_R=0.5$ s and $h(t)$ of Eq. (5). An $e_n^2(t)$ and an $n(t)$ of Eq. (6) with an SNR of 3 dB are shown in Figs. 3(e) and (f), and we show $e_y^2(t)$ ($=e_x^2(t)*e_h^2(t)+e_n^2(t)$ and the observed noisy reverberant signal $y(t)$ ($=x(t)*h(t)+n(t)$) in Figs. 3(g) and (h). The left panels ((a), (c), (e), and (g)) show the power envelopes and the right panels ((b), (d), (f), and (h)) show the corresponding signals. As shown in this figure, $m(f_m)$ decreased from 1.0 (in Fig. 3(a)) to 0.404×0.5 (the maximum deviation of the envelope between the dotted lines in Fig. 3(e) relative to that in Fig. 3(a) and the reduction in Fig. 3(g)). The solid line in Fig. 3(g) shows the restored power envelope $\hat{e}_x^2(t)$ obtained from the noisy reverberant power envelope $e_y^2(t)$ (Fig. 3(g)) using Eqs. (16) and (17) with $T_R=0.5$ s and SNR = 3 dB. It is shown that using power envelope restoration can precisely restore the power envelope from a noisy reverberant signal in terms of its shape and magnitude.

4. EVALUATION

We now describe how we carried out the following simulations to evaluate the proposed method. The speech signals were three Japanese sentences (/aikawarazu/, /shinbun/, /joudan/) uttered by ten speakers (five males and five females) from the ATR database [11]. We used 100 artificial impulse responses $h(t)$ and 100 white noise signals $n(t)$. Two reverberation times ($T_R=0.5$ and 2.0 s) were used. Signal to noise ratios (SNRs) between $x(t)$ and $n(t)$ were fixed at 10 and 0 dB. All reverberant signals ($6,000=10\times3\times2\times100$) and all noisy signals ($6,000=10\times3\times2\times100$) were generated by convolving $x(t)$ with $h(t)$ and by adding $n(t)$ to $x(t)$. All noisy reverberant signals $y(t)$ ($12,000=10\times3\times2\times100$) were also used. The sampling frequency of signal f_s is 20 kHz. We used a filterbank for speech restoration, and divided the signal into 100 channels (100 Hz bandwidth).

In this paper, to evaluate both the error and similarity between temporal power envelopes (with magnitude and shape), we then used correlation (Corr) and SNR as evaluation measures. Improvements in these measures, therefore, show the extent to which using our method improve accuracy of the restoration. These measures are defined as

$$\mathrm{Corr}(e_x^2,\hat{e}_x^2)=\frac{\int_0^T\left(e_x^2(t)-\overline{e_x^2}\right)\left(\hat{e}_x^2(t)-\overline{\hat{e}_x^2}\right)dt}{\sqrt{\left\{\int_0^T(e_x^2(t)-\overline{e_x^2})dt\right\}\left\{\int_0^T(\hat{e}_x^2(t)-\overline{\hat{e}_x^2})dt\right\}}}, \quad (21)$$

$$\mathrm{SNR}(e_x^2,\hat{e}_x^2)=10\log_{10}\frac{\int_0^T(e_x^2(t))^2dt}{\int_0^T(e_x^2(t)-\hat{e}_x^2(t))^2dt}, \quad (22)$$

where $\overline{e_x^2}$ is the average value of $e_x^2(t)$, $\hat{e}_x^2(t)$ is the restored temporal power envelope. The improvements in Corr and SNR are calculated from $\mathrm{Corr}(e_x^2,\hat{e}_x^2)-\mathrm{Corr}(e_x^2,e_y^2)$, and

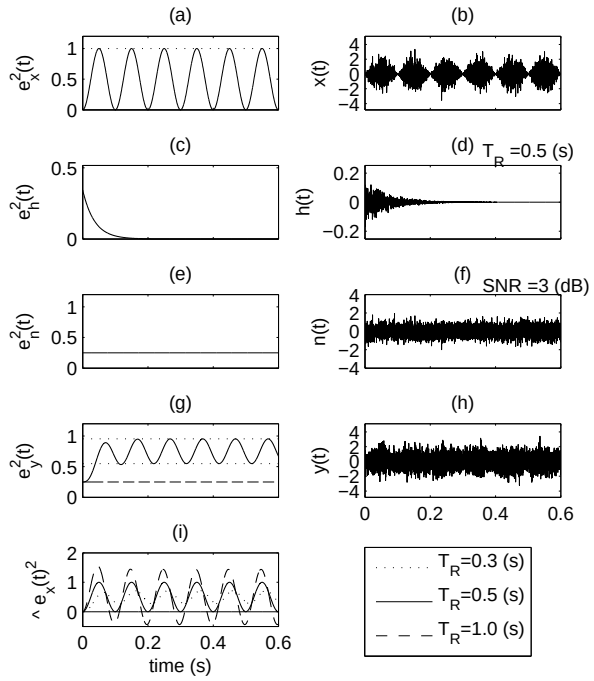

Figure 3: Example of relationship between power envelopes of system based on the MTF concept: (a) power envelope $e_x^2(t)$ of (b) original signal $x(t)$, (c) power envelope $e_h^2(t)$ of (d) simulated room impulse response $h(t)$ ($T_R = 0.5$ s), (e) power envelope $e_n^2(t)$ of (f) noise signal $n(t)$, (g) power envelope $e_y^2(t)$ derived from $e_x^2(t) * e_h^2(t) + e_n^2(t)$, (h) noisy reverberant signal $y(t)$ derived from $x(t) * h(t) + n(t)$, and (i) restored power envelope $\hat{e}_x^2(t)$.

$\mathrm{SNR}(e_x^2, \hat{e}_x^2) - \mathrm{SNR}(e_x^2, e_y^2)$. Note that the positive values indicate the temporal power envelope and waveform of speech were restored from the noisy signal to a certain degree.

The improvement of Corr and SNR in each channel under the reverberant and noisy conditions are shown in Figs. 4 and 5. The height of bar shows the mean value. Although the error bar (standard deviation) was not plotted in here for clearly viewing, note that the standard deviation at each condition was not too much. The improvement in Corr was almost zero in the noisy conditions while the improvement in Corr was too much in the reverberant conditions. The improvements in the average SNR increased as the SNRs decreased in the noisy conditions while the improvement in SNR was almost constant in the reverberant conditions. These results show that using the proposed method can improve the temporal power envelope of the input signal from the reverberant or noisy signal. In particular, the proposed method greatly improved Corr in the reverberant conditions while the SNR in the noisy conditions was also significantly improved.

The improvement in Corr and improved SNR in each channel under the noisy reverberant conditions are shown in Fig. 6. These improvements show that the proposed method can be used to improve the temporal power envelope from the noisy reverberant signals to an adequate level. Furthermore, doing this can simultaneously suppress the effects of reverberation and noise.

An example of a restoration when we used the proposed method for Japanese sentence (/aikawarazu/) of a male speaker (Mau) with $T_R = 0.5$ s and SNR $= 10$ dB is shown in Fig. 7. The power envelopes of only a quarter of all channels are plotted in this figure (#1, #5, #9, and so on). These magnitudes were normalized in each channel to view matches between $e_x^2(t)$s and $\hat{e}_x^2(t)$s. We can see many matches between $e_x^2(t)$s and the restored $\hat{e}_x^2(t)$s in panel (d), but there are fewer matches in panel (c). These results demonstrate that the proposed method can be used to adequately restore the power envelope from noisy reverberant speech $y(t)$.

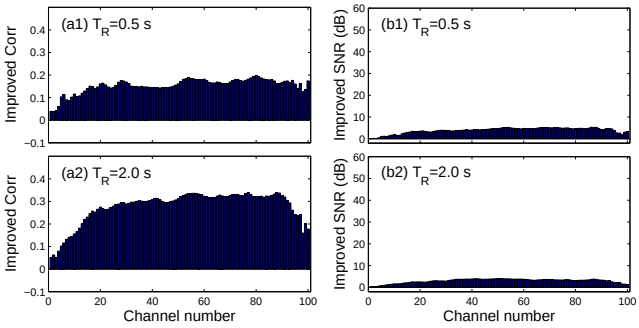

Figure 4: Improvement in dereverberation accuracy for the power envelope of speech on the filterbank: (a) improved correlations and (b) improved SNRs. $T_R = 0.5$ and 2.0 s.

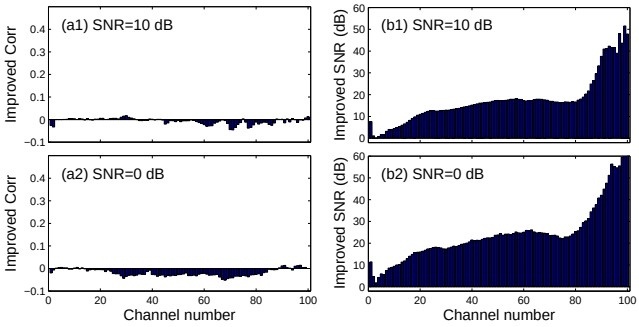

Figure 5: Improvement in suppression accuracy for the power envelope of speech on the filterbank: (a) improved correlations and (b) improved SNRs. SNR $= 10$ and 0 dB.

5. CONCLUSION

We have explained how the MTF concept helps to suppress the effects of reverberation and noise for improving the intelligibility of speech. We then proposed a power envelope restoration method that is based on the MTF concept. We have carried out simulations to evaluate whether using the proposed method can restore smearing of the temporal power envelope of speech signals in both noisy and reverberant environments. We found that using the proposed method can be used to simultaneously restore the temporal power envelopes and to suppress the noise and reverberation effects.

In our future work, we will (1) reconsider an adaptive restoration method on the time-frequency divisions for restoring noisy reverberant speech, (2) attempt to construct a way of restoring carriers in the filterbank model to resynthesize the restored speech, and (3) test whether our approach can suppress the reduction caused by reverberation and noise

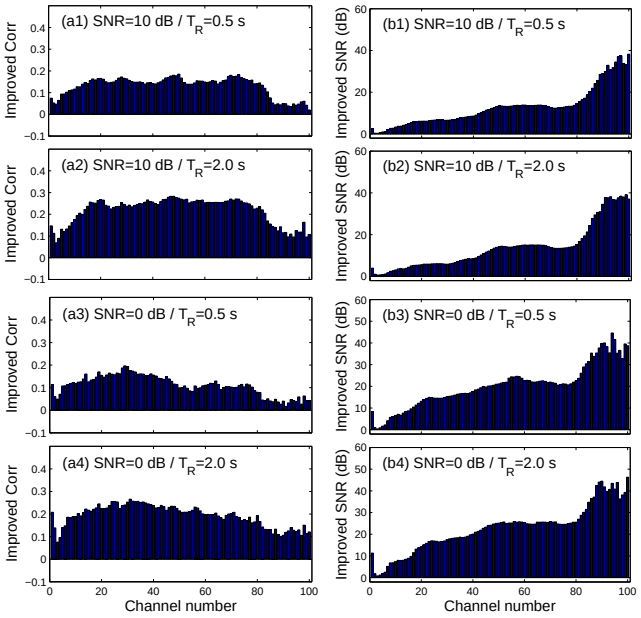

Figure 6: Improvement in restoration accuracy for the power envelope of speech on the filterbank: (a) improved Corrs and (b) improved SNRs. $T_R = 0.5$ and 2.0 s. SNR $= 10$ and 0 dB.

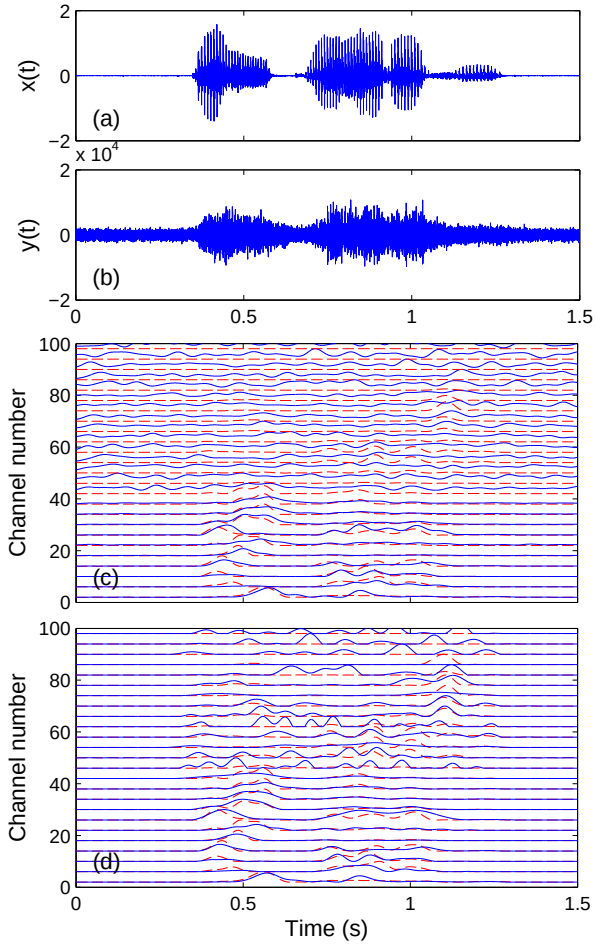

Figure 7: Example of the power envelope restoration: (a) original speech $x(t)$, (b) noisy reverberant speech $y(t)$ in the case of $T_R = 0.5$ s and SNR $= 10$ dB, (c) no processing ($e_y^2(t)$s; solid lines) and (d) restoration using proposed method ($\hat{e}_x^2(t)$s; solid lines). Dashed lines show $e_x^2(t)$s.

to the intelligibility of speech.

Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (No. 18680017) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan and the Strategic Information and COmmunications R&D Promotion ProgrammE (SCOPE) (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP.*, **27**(2), 113–120, 1979.

[2] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *ICASSP'87*, **1**, 177–180, 1987.

[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J.Acoust. Soc. Am.*, **66**(1), 166–169, July 1979.

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP.*, Vol. **36**(2), 145–152, Feb. 1988

[5] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech enhancement in noisy reverberant environment," *Proc. Interspeech-2007.*, 854–857, Aug. 2007.

[6] T. Houtgast and H. J. M. Steeneken, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," *Acustica.*, **28**, 66–73, 1973.

[7] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. & Tech.* **25**(4), 232–242, 2004.

[8] M. Unoki, K. Sakata, M. Furkawa, and M. Akagi, "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. & Tech.*, **25**(4), 243–254. 2004.

[9] M. Unoki, M. Toi, and M. Akagi, "Development of the MTF-based speech dereverberation method using adaptive time-frequency division," *Proc. Forum Acusticum2005*, 51–56, Budapest, 2005.

[10] X. Lu, M. Unoki, and M. Akagi, "Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoust. Sci. & Tech.*, **29**(6), 351–361, 2008.

[11] T. Takeda *et al.*, Speech Database User's Manual,ATR Tecnical Report, TR-I-0028, 1988.

[12] D. Ying, Y. Shi, X. Lu, J. Dang, and F. Soong, "Robust voice activity detection based on noise eigenspace," *Acout. Sci. & Tech.*, **28**(6), 413–423, 2007.