Journal of Signal Processing, Vol. 13, No. 4, pp. 335-338, July 2009

SELECTED PAPER

Study on a Method of Suppressing Noise Based on the MTF Concept

Yutaka Yamasaki and Masashi Unoki

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292, Japan
Phone/FAX:+81-761-51-1699 (Ex. 1391) /+81-761-51-1149
E-mail: {yutaka1017, unoki}@jaist.ac.jp

Abstract

Many methods of enhancing speech have recently been proposed to suppress the effects of noise or reverberation. Although most of these have aimed to only enhance noisy speech or reverberant speech, they have not simultaneously been able to enhance noisy reverberant speech. As MTF can be used to predict the loss of speech intelligibility due to noise and reverberation, it may be possible to simultaneously suppress the effects of deterioration due to noise and reverberation, by utilizing MTF-based processing. No methods of suppressing noise based on the MTF concept have yet been considered while a method of dereverberation based on the MTF concept has already been proposed. We propose a method of restoring temporal power envelopes from noisy speech based on MTF. The proposed method suppresses noise by restoring smeared MTF. We carried out simulations on suppressing noise in noisy speech to objectively evaluate the model we propose. The results obtained from evaluating the method demonstrated that the proposed approach could effectively suppress noise.

1. Introduction

Significant features of speech are smeared in real environments due to noise and reverberation so that the sound quality and intelligibility of observed speech are drastically reduced. Noisy reverberant speech (noise suppression and dereverberation) therefore needs to be enhanced in various speech-signal processes, such as those in hearing-aid systems and preprocessing in automatic speech recognition systems.

There are several well-known methods of suppression that can be used to remove the effects of noise or reverberation in either noisy or reverberant environments. There have been, for example, the spectral subtraction proposed by Boll [1], the Kalman filtering proposed by Paliwal and Basu [2], the minimum-phase inverse filtering method proposed by Neely and Allen [3], and the multiple input/output inverse theorem (MINT) proposed by Miyoshi and Kaneda [4]. Although these methods can work well in either noisy or reverberant environments, they cannot work simultaneously in both noisy and reverberant environments. Kinoshita *et al.* recently studied a strategy of enhancing speech in noisy reverberant environments, by taking into consideration two sequential processes, i.e., noise reduction using spectral subtraction for noisy reverberant speech and then dereverberation using linear prediction for noise-reduced reverberant speech [5]. However, Kinoshita's modeling seems to be too complex. We thought that the best solution would be able to simultaneously deal with both additive noise and reverberant effects.

However, Houtgast and Steeneken proposed a method of prediction that could assess the effects of an enclosure on speech intelligibility in both noisy and reverberant environments by using the modulation transfer function (MTF) [6]. The MTF concept made it possible to simultaneously suppress both noise and reverberation.

Unoki *et al.* proposed a method of inverse filtering for temporal power envelopes based on the MTF concept [7, 8, 9]. Their method assumed environments that had reverberation and it improved speech intelligibility in these by about 30%. We propose a method of speech enhancement based on MTF that can simultaneously suppress noise and reverberation, so that it can improve speech intelligibility lost through additive noise and reverberation.

Our goal was to propose a method of speech enhancement to reduce noise and dereverberation. This paper proposes a method of suppressing noise by restoring smeared MTF.

2. MTF Concept

The MTF concept was proposed by Houtgast and Steeneken to account for the relation between the degree of modulation in the envelopes of input and output signals and the characteristics of the enclosure. This concept was introduced as a measure in room acoustics to assess what effect the enclosure had on speech intelligibility. The input and output temporal power envelopes in their concept are defined as

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

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

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. Here, $m(f_m)$ is the modulation index of the modulation frequency and is referred as to MTF.

2.1 MTF in noisy environments

This section explains MTF in noisy environments. The input temporal power envelope, $e_x^2(t)$, is defined as

$$e_x^2(t) = \overline{e_x^2}(1 + \cos(2\pi f_m t)) \quad (3)$$

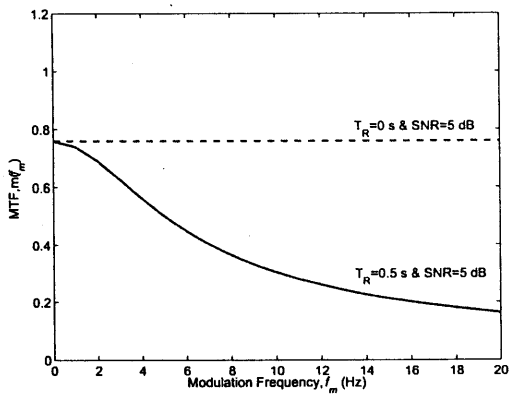

Figure 1: Theoretical representation of MTF, $m(f_m)$, in noisy and reverberant environments

Under additive noise conditions, the output temporal power envelope, $e_y^2(t)$, is defined as

$$\begin{aligned} e_y^2(t) &= \overline{e_x^2}\{1+\cos(2\pi f_m t)\} + e_n^2(t) \\ &= \left(\overline{e_x^2}+\overline{e_n^2}\right)\{1+m(f_m)\cos(2\pi f_m t)\} \end{aligned} \quad (4)$$

where $e_n^2(t)$ is the temporal power envelope of the noise signal. $\overline{e_n^2} = \frac{1}{T}\int_0^T e_n^2(t)dt$ because $e_n^2(t)$ is assumed to be constant over the time. Here T is the duration of the signal. The complex MTF in noisy environments is defined as

$$m(f_m) = \frac{\overline{e_x^2}}{\overline{e_x^2}+\overline{e_n^2}} = \frac{1}{1+10^{-(\mathrm{SNR})/10}} \quad (5)$$

where the signal to noise ration (SNR) equals $10\log_{10}(\overline{e_x^2}/\overline{e_n^2})$ in dB in decibels. This MTF is independent of modulation frequency f_m. With an SNR of 5 dB, $m(f_m)$ is about 0.76 (as plotted in Fig. 1).

2.2 MTF in noisy and reverberant environments

The MTF in reverberant environments, $m(f_m)$, can be represented as [6, 7, 8, 9]

$$m(f_m) = \left[1+\left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-1/2} \quad (6)$$

where T_R is the reverberation time. The MTF in reverberant environments depends on f_m. This equation indicates low-pass characteristics. The MTF in noisy and reverberation environments calculated from Eqs. (1), (2), (5), and (6) can be represented as

$$m(f_m) = \left[1+\left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-1/2} \times \left(1+10^{-(\mathrm{SNR})/10}\right)^{-1} \quad (7)$$

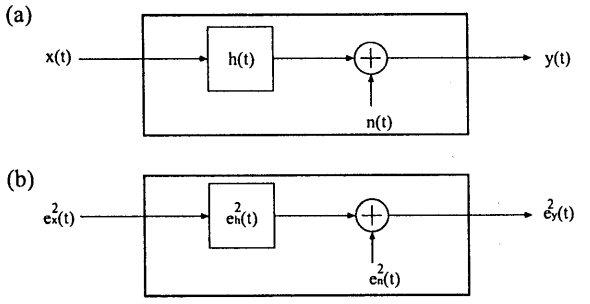

Figure 2: Diagram of transfer function: (a) is for signal, and (b) is for temporal power envelope.

Hence, noise and reverberant can be suppressed when using the inverse filtering of the MTF in Eq. (7).

3. Method of Suppressing Noise

3.1 Model concept based on MTF

The output signal, the input signal, the impulse response, and the noise signal have been assumed to correspond to $y(t)$, $x(t)$, $h(t)$, and $n(t)$ in this paper, as outlined in Fig. 2(a). These four were modeled based on the MTF concept as

$$y(t) = h(t) * x(t) + n(t) \quad (8)$$
$$h(t) = e_h(t)c_h(t) \quad (9)$$
$$x(t) = e_x(t)c_x(t) \quad (10)$$
$$n(t) = e_n(t)c_n(t) \quad (11)$$
$$\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau) \quad (12)$$

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 variables. $\langle\cdot\rangle$ is an ensemble average operation. In this model, $e_y^2(t)$ can be derived as

$$\langle y^2(t)\rangle = \langle h^2(t) * x^2(t)\rangle + \langle n^2(t)\rangle \quad (13)$$
$$e_y^2(t) = e_h^2(t) * e_x^2(t) + e_n^2(t) \quad (14)$$

(see [7, 8, 9] for a detailed derivation of Eq. (13).) The relation of temporal power envelopes was used in our study. However, only noisy environments have been considered in this paper for our proposed method of suppressing MTF-based noise.

3.2 Extraction of temporal power envelopes

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

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

where LPF[·] is low-pass filtering. Hilbert(·) is the Hilbert transform. This method is based on calculating the instantaneous amplitude of the signal, and low-pass filtering is used

as post-processing to remove the higher frequency components in the power envelopes. We used LPF with a cut-off frequency of 20 Hz. Unoki *et al.*'s method was adopted to extract the temporal power envelopes [7, 8, 9].

3.3 Implementation

This section explains the method of suppressing noise based on the MTF concept. The modulation index and the averaged power in Eq. (4) are affected by noise. We restore the averaged power levels to suppress the noise effects. Eq. (16) is the offset value of the averaged power given by

$$\mathbf{OV} = \frac{\overline{e_x^2}}{\overline{e_x^2} + \overline{e_n^2}} \tag{16}$$

By substituting Eq. (16) into Eq. (4), we can obtain

$$\overline{e_x^2} + \overline{e_x^2} \cdot m(f_m) \cdot \cos(2\pi f_m t) \tag{17}$$

By multiplying the second term of Eq. (17) by $1/m(f_m)$ to restore the modulation index, we can obtain

$$\begin{aligned} \hat{e}_x^2(t) &= \overline{e_x^2} + \left(\overline{e_x^2} \cdot m(f_m) \cdot \cos 2\pi f_m t\right) \times \frac{1}{m(f_m)} \\ &= \overline{e_x^2}(1 + \cos(2\pi f_m t)) \end{aligned} \tag{18}$$

This obtains the noise-suppressed temporal power envelope, $\hat{e}_x^2(t)$. This algorithm is the method of suppressing noise based on the MTF concept.

The proposed approach is equal to the method of subtracting the average value of noise temporal power envelopes from the output temporal power envelope. Since, the proposed method is based on the MTF concept, it should be able to deal with not only additive noise but also reverberation. Therefore, if the proposed approach is incorporated into MTF-based dereverberation, the combined method can simultaneously suppress both noise and reverberation.

$\overline{e_x^2}$ and $\overline{e_n^2}$ are needed to calculate the MTF in Eq. (5). $\overline{e_n^2}$ is estimated from the non-speech sections. $\overline{e_x^2}$ is estimated by subtracting $\overline{e_n^2}$ from $\overline{e_y^2}$. The MTF in Eq. (18) can be calculated by using this SNR.

4. Evaluation

We carried out the following simulations to evaluate the proposed model. The speech signals were three Japanese sentences (/aikawarazu/, /shinbun/, /joudan/) uttered by ten speakers (five males and five females) from the ATR database [10]. We used 100 white noise signals, $n(t)$. The SNRs were fixed at 20, 10, 5, 0, and -5 dB. All noisy signals ($15,000 = 10 \times 3 \times 5 \times 100$) were generated by adding $x(t)$ to $n(t)$. The sampling frequency of the signal was 20 kHz. We used a filterbank for speech restoration, and divided the signal into 100 bands. The bandwidth of all channels was set to 100 Hz.

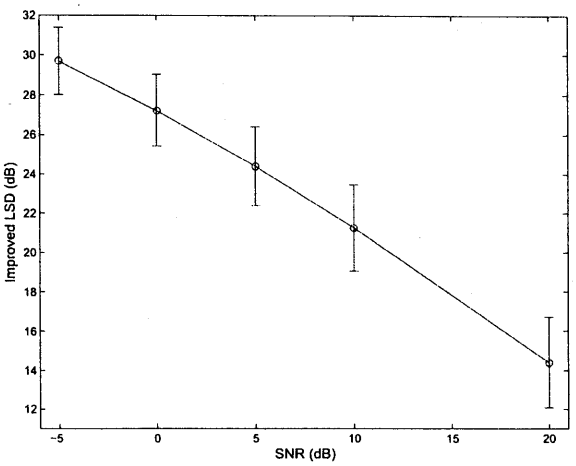

Figure 3: Mean improvement in restoration accuracy (LSD): Error bars represent standard deviation.

The correlation (Corr), SNR, and log spectrum distance (LSD) were used as measures in these simulations to evaluate improvements in the accuracy of restoration achieved with our method, These measures are defined as

$$\begin{aligned} &\mathrm{Corr}(e_x^2, \hat{e}_x^2) \\ &= \frac{\int_0^T \left(e_x^2(t) - \overline{e_x^2}\right)\left(\hat{e}_x^2(t) - \overline{\hat{e}_x^2}\right) dt}{\sqrt{\left\{\int_0^T (e_x^2(t) - \overline{e_x^2(t)})dt\right\}\left\{\int_0^T (\hat{e}_x^2(t) - \overline{\hat{e}_x^2(t)})dt\right\}}} \end{aligned} \tag{19}$$

$$\mathrm{SNR}(e_x^2, \hat{e}_x^2) = 10\log_{10}\frac{\int_0^T (e_x^2(t))^2 dt}{\int_0^T (e_x^2(t) - \hat{e}_x^2(t))^2 dt}, \tag{20}$$

$$\mathrm{LSD}(S_x, \hat{S}_x) = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log_{10}\frac{|S_x(\omega)|}{|\hat{S}_x(\omega)|}\right)^2} \tag{21}$$

where $\overline{e_x^2}$ is the average value of $e_x^2(t)$, and $\hat{e}_x^2(t)$ is the restored temporal power envelope. W is the upper frequency (here, it is 10 kHz), and $S_x(\omega)$ and $\hat{S}_x(\omega)$ are the amplitude spectra of $x(t)$ and $y(t)$ or $\hat{x}(t)$. The improvements in Corr, SNR, and LSD are calculated from $\mathrm{Corr}(e_x^2, \hat{e}_x^2) - \mathrm{Corr}(e_x^2, e_y^2)$, $\mathrm{SNR}(e_x^2, \hat{e}_x^2) - \mathrm{SNR}(e_x^2, e_y^2)$, and $\mathrm{LSD}(S_x(\omega), S_y(\omega)) - \mathrm{LSD}(S_x(\omega), \hat{S}_x(\omega))$. Note that positive values indicate the temporal power envelopes and waveforms of speech were restored from noisy signals to a certain degree.

Figure 3 plots the improvement in LSD. The maximum improvement in LSD was about 30 dB. Figure 4 shows the improvement in Corr and the improvement SNR in all channels. The heights of the bars and error bars correspond to the mean and standard deviation. The improvement in Corr was constant. The improvements in SNR increased as SNRs decreased. These results indicate that the proposed method can

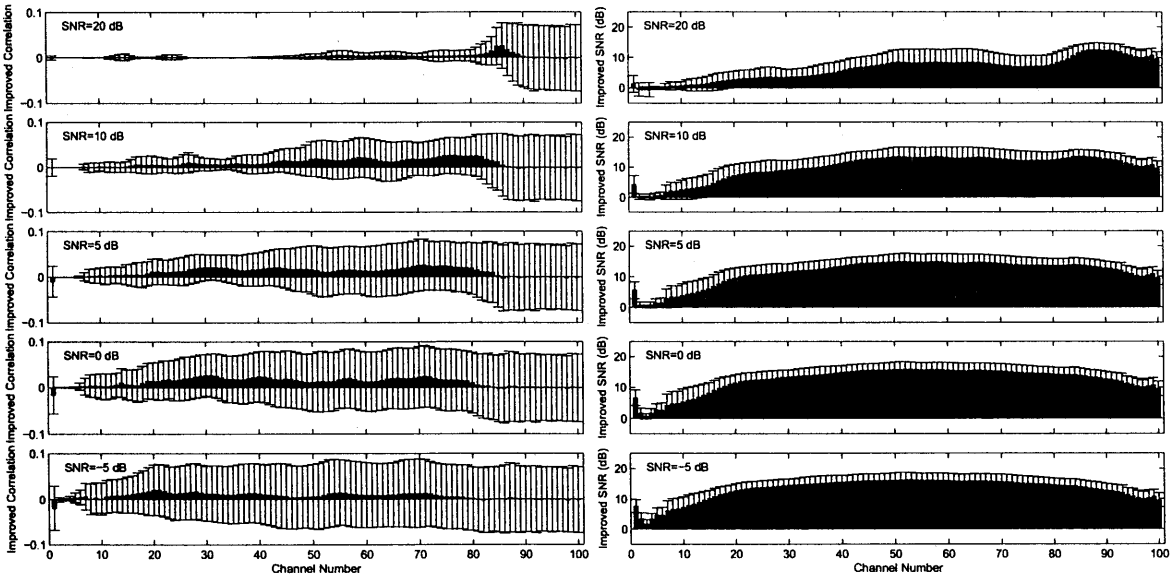

Figure 4: Improved accuracy for temporal power envelopes of speech in filterbank

improve temporal power envelopes and the waveforms of the input signals from the noisy signals.

5. Conclusion and Future Perspectives

We introduced the MTF concept and proposed a method of suppressing noise based on the MTF concept by restoring smeared MTF. We carried out simulations in which the proposed approach was applied to temporal power envelopes to restore 15,000 noisy speech signals. We found that the proposed method could be used to adequately restore the temporal power envelopes and to suppress the noise effects of noisy signals.

We intend to further improve the method, consider carrier restoration, and carry out subjective evaluations in future work. We further intend to propose a method of suppressing noise and reverberation by restoring the smeared MTF in noisy and reverberant environments using the inverse filtering of MTF in Eq. (7).

Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (No. 18680017) made available by the Ministry of Education, Culture, Sports, Science, and Technology, Japan. It was also partially supported by 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., Vol. 27, No. 2, pp. 113–120, 1979.

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

[3] S. T. Neely, J. B. Allen: Invertibility of a room impulse response, J.Acoust. Soc. Am., Vol. 66, No. 1, pp. 166–169, 1979.

[4] M. Miyoshi, Y. Kaneda: Inverse filtering of room acoustics, IEEE Trans. ASSP., Vol. 36, No. 2, pp. 145–152, 1988.

[5] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi: Multi-step linear prediction based speech enhancement in noisy reverberant environment, Proc. Interspeech2007., pp. 854–857, 2007.

[6] T. Houtgast and H. J. M. Steeneken: The Modulation transfer function in room acoustics as a predictor of speech intelligibility, Acustica, Vol. 28, pp. 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., Vol. 25, No. 4, pp. 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., Vol. 25, No. 4, pp. 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 in Budapest, pp. 51–56, 2005.

[10] T. Takeda et al.: Speech Database User's Manual,ATR Technical Report, TR-I-0028, 1988.