MTF-based Sub-band Power-envelope Restoration for Robust Speech Recognitionin Noisy Reverberant Environments

Shota Morita[1)], Xugang Lu[2)], Masashi Unoki[1)], Masato Akagi[1)], Rüdiger Hoffmann[3)]
[1)] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
E-mail: s-morita, unoki, akagi@jaist.ac.jp
[2)] National Institute of Information and Communications Technology, Japan
E-mail: xugang.lu@nict.go.jp
[3)] Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
E-mail: Ruediger.Hoffmann@ias.et.tu-dresden.de

***Abstract*—Many speech enhancement methods have been proposed to suppress the effect of either background noise or reverberation for automatic speech recognition (ASR) systems. However, most of these methods cannot simultaneously reduce the effects of both, and no method reduces the effects of both in a unified strategy for ASR systems in noisy reverberant environments. We previously proposed a method for restoring the speech power envelope from noisy reverberant speech based on a simple modulation transfer function (MTF) concept. The method does not require the impulse response and noise conditions of the room acoustics to be measured. In this study, we further tested the proposed method as a front-end for ASR systems in noisy reverberant environments. Noisy reverberant speech signals were obtained by adding white noise to reverberant speech produced by convoluting clean speech signals (from the AURORA-2J, a continuous Japanese digit speech) with artificially-made impulse response of room acoustics. The recognition performance based on the conventional Mel frequency cepstral coefficient feature was used as a baseline. Compared with the baseline, the proposed method obtained 12.19 % relative improvement in the error reduction rate (averaged of all tested noisy reverberant environments).**

I. Introduction

The speech recognition rate of automatic speech recognition (ASR) systems is drastically reduced in real environments due to degradation of the speech features caused by reverberation and background noise. Achieving robust speech recognition in a noisy reverberant environment is therefore an important issue. Several well-known suppression methods are used to remove the effects of either background noise or reverberation, e.g., the spectral subtraction method [1], minimum-phase inverse filtering method [2], and RASTA filtering [3]. These methods work well in either noisy or reverberant environments, but they do not work well when background noise and reverberation exist simultaneously in the environment.

Kinoshita *et al*. have proposed a method to enhance speech recorded in a noisy reverberant environment. Two sequential processes were considered in their method: noise reduction using spectral subtraction, and then dereverberation using linear prediction for noise-reduced reverberant speech [4]. This method could restore the spectrogram of noisy reverberant speech with consideration of the different effects of background noise and reverberation. On the other hand, Houtgast and Steeneken have proposed a method to predict the speech intelligibility in an enclosure as a noisy and reverberant environment, and they unified the effects by using the modulation transfer function (MTF) concept [5]. Unoki *et al*. proposed the use of a power envelope inverse filtering method based on the MTF concept [6]; they obtained 30% relative improvement in the error reduction rate for ASR in reverberant environments [7]. Recently, the power envelope restoration method based on the MTF concept was proposed for noisy reverberant speech [8]. However, it is not clarified the performance of the method as applicative system in noisy and/or reverberant environments. In this study, we applied the method as a front-end processor for ASR, therefore we investigate the performance of the method to clarify its effectiveness in noisy and/or reverberant environments.

II. MTF-based modeling in noisy reverberant environments

A. Modulation transfer function concept

The MTF concept was proposed by Houtgast and Steeneken to predict speech intelligibility in room acoustics [5]. It is used as a modulation index accounting for the relationship of the degree of modulation of the temporal envelopes between input and output signals in an enclosure. The input and output temporal power envelopes are defined as

$$\text{Input:} \quad \overline{I_i^2}(1+\cos(2\pi f_m t)) \tag{1}$$

$$\text{Output:} \quad \overline{I_o^2}\{1+m(f_m)\cos(2\pi f_m(t-\theta))\}, \tag{2}$$

where $\overline{I_i^2}$ and $\overline{I_o^2}$ are the input and output intensities, f_m is the modulation frequency, and θ is the phase information. $m(f_m)$ is the modulation index of the temporal power envelope that is referred to as MTF.

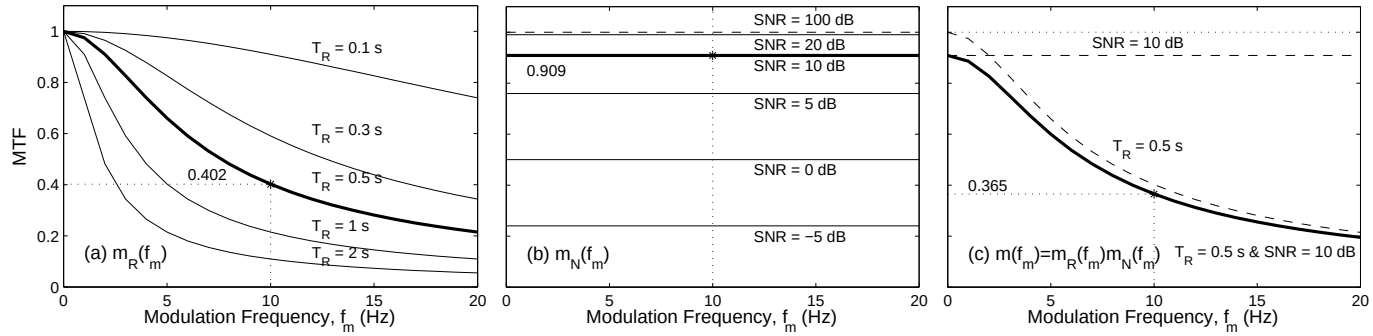

Fig. 1. Theoretical representation of MTF, $m(f_m)$ in both noisy and/or reverberant environments. Bold solid line indicates MTF with T_R=0.5 s and SNR = 10 dB.

B. Signal modeling based on MTF

We assume the input, output, impulse response, and noise signal to be $x(t)$, $y(t)$, $h(t)$, and $n(t)$, respectively. They are modeled based on the MTF as

$$\begin{aligned} y(t) &= h(t) * x(t) + n(t) && (3)\\ x(t) &= e_x(t)c_x(t) && (4)\\ h(t) &= e_h(t)c_h(t) = a\exp(-6.9t/T_R)c_h(t) && (5)\\ n(t) &= e_n(t)c_n(t), && (6)\end{aligned}$$

where $e_x(t)$, $e_h(t)$, and $e_n(t)$ are the temporal envelopes of $x(t)$, $h(t)$, and $n(t)$ respectively. $c_x(t)$, $c_h(t)$, and $c_n(t)$ are the carrier of $x(t)$, $h(t)$, and $n(t)$ (a Gaussian white noise) respectively. T_R is the reverberation time. Based on stochastic analysis, the following is derived [6]:

$$\begin{aligned} \langle y^2(t)\rangle &= \langle h^2(t) * x^2(t)\rangle + \langle n^2(t)\rangle && (7)\\ e_y^2(t) &= e_h^2(t) * e_x^2(t) + e_n^2(t). && (8)\end{aligned}$$

In the derivation, $\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau)$ with $c_l \in \{c_x, c_h, c_n\}$, and $\langle\cdot\rangle$ is an ensemble average operation.

C. MTF in noisy and/or reverberant environments

The complex MTF in reverberant environment is defined as

$$m_R(f_m) = \left|\frac{\beta}{\alpha}\right| = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-1/2}, \quad (9)$$

where $\alpha = \int_0^\infty h^2(t)dt$ and $\beta = \int_0^\infty h^2(t)\exp(-j\omega_m t)dt$. f_m is the modulation frequency and T_R is the reverberation time. The theoretical analysis is shown in Fig. 1(a). The complex MTF in noisy environments is defined as

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

where $\mathrm{SNR} = 10\log_{10}(\overline{e_x^2}/\overline{e_n^2})$ in dB. The theoretical analysis is shown in Fig. 1(b). The MTF in noisy reverberant environments can be represented as

$$\begin{aligned} m(f_m) &= m_R(f_m)\cdot m_N(f_m)\\ &= \frac{1}{\sqrt{1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2}\left(1 + 10^{-\frac{\mathrm{SNR}}{10}}\right)}. \quad (11)\end{aligned}$$

The MTF in noisy reverberant environments depends on f_m, T_R, and SNR. It means the low-pass characteristics result from reverberation as a function of T_R and the constant attenuation results from noise as a function of SNR. An example is shown in Fig. 1(c). In this figure, $T_R = 0.5$ s, SNR = 10 dB, and $m(f_m)$ at $f_m = 10$ Hz is 0.365 (= 0.402×0.909). Hence, the effect of noise and reverberation can be suppressed by using the inverse filtering of MTF formulated in Eq. (11).

III. Sub-band power envelope restoration based on MTF

We previously explained the MTF-based sub-band power envelope restoration method [8]. A block-diagram of the method is shown in Fig. 2. It consists of (i) power envelope extraction, (ii) power envelope subtraction, and (iii) power envelope inverse filtering with parameter estimation. The constant bandwidth filterbank is used in the signal analysis. The sub-band power envelope $e_y^2(t)$ is extracted by

$$e_y^2(t) = \textbf{LPF}\left[|y(t) + j\textbf{Hilbert}(y(t))|^2\right], \quad (12)$$

where Hilbert$(\cdot)$ is the Hilbert transform and LPF$[\cdot]$ is low-pass filtering with a cut-off frequency of 20 Hz [6].

The power envelope subtraction is used to suppress the additive noise effect. The first term in Eq. (8) is estimated as

$$\begin{aligned} \hat{e}_x^2(t) &= \overline{e_x^2}\left(1 + m_N(f_m)\cos(2\pi f_m t) \times \frac{1}{m_N(f_m)}\right)\\ &= e_y^2 - \overline{e_n^2}. \quad (13)\end{aligned}$$

In the estimation, a robust VAD algorithm is used to calculate the average power of noise $\overline{e_n^2}$ from the observed $e_y^2(t)$.

On the basis of these results, e_x^2 can be recovered by inverse filtering $\hat{e}_y^2(t) = e_x^2(t) * e_h^2(t)$ in Eq.(8) with $e_h^2(t)$. The transfer functions of power envelope $E_x(z)$, $E_h(z)$, and $E_n(z)$ are assumed to be the z-transforms of $e_x^2(n)$, $e_h^2(n)$, and $e_n^2(n)$, respectively. Thus, $E_x(z)$ can be determined as

$$E_x(z) = \frac{E_x(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\}, \quad (14)$$

where f_s is the sampling frequency. The power envelope $e_x^2(n)$ can be obtained from the inverse z-transform of $E_x(z)$. Two

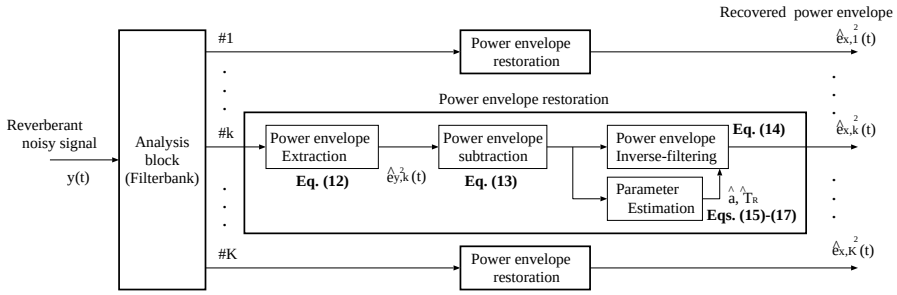

Fig. 2. MTF-based sub-band power envelope restoration method.

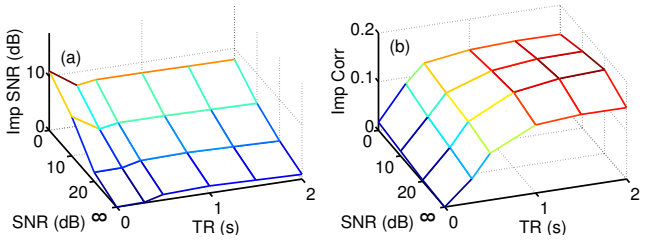

Fig. 3. Improvement of restored power envelopes in noisy reverberant environments: (a) improved SNR, (b) improved Corr.

parameters in Eq. (14) are estimated as follows [6]:

$$\hat{T}_R = \underset{0\le T_R\le T_{R,\max}}{\arg\min}\left\{\frac{dT_P(T_R)}{dT_R}\right\}, \quad (15)$$

$$T_P(T_R) = \min\left(\underset{t_{\min}\le t\le t_{\max}}{\arg\min}\left|\hat{e}_{x,n,T_R}(t)^2-\theta\right|\right), \quad (16)$$

$$\hat{a} = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt}. \quad (17)$$

IV. Evaluations

A. Evaluation of power envelope restoration

Simulations were carried out to evaluate the proposed method in artificial noisy reverberant environments, because to make sure of processing based on the basic principle of the method. We used 100 clean speech sentences from the AURORA-2J, a continuous Japanese digit speech (denoted as $x(t)$). 10 artificial impulse responses of room acoustics (denoted as $h(t)$) and 10 Gaussian white noise signals were used to make the noisy reverberant speech. Five reverberation conditions were simulated with reverberation time (T_R) of 0.3, 0.5, 1.0, 1.5, and 2.0 s. Signal to noise ratios (SNRs) between $x(t)$ and $n(t)$ were fixed at $20, 10$, and 0 dB. All reverberant signals ($5,000 = 100\times5\times10$) were generated by convoluting $x(t)$ with $h(t)$. All noisy reverberant signals $y(t)(15,000 = 100\times5\times3\times10)$ were generated by convoluting $x(t)$ with $h(t)$ and adding $n(t)$. The sampling frequency of signal f_s was 8 kHz. 40 sub-band filters (100 Hz band width) were used for signal decomposition.

To measure the accuracy of the speech power envelope restoration, we used (i) SNR (ratio between power envelope of original signal and restored power envelope as SNR($e_x^2, \hat{e}_x^2$)), and (ii) Corr. (correlation between power envelope of original signal and restored power envelope as Corr($e_x^2, \hat{e}_x^2$)). The improvements are calculated as SNR($e_x^2, \hat{e_x^2}$)$-$SNR(e_x^2, e_y^2) and Corr($e_x^2, \hat{e_x^2}$)$-$Corr(e_x^2, e_y^2), respectively.

The average improvements in SNR and Corr. are shown in Figs. 3(a) and (b), respectively. From these figures, the trend in a reverberant environment (SNR = ∞), which is the improvement in Corr., increased as reverberation time increased, and the trend in noisy environment (T_R = 0), which is the improvement in SNR, increased as the power of additive noise increased. We can see that the trend in a noisy reverberant environment is a combination of the trends in noisy environments and reverberant environments. Thus, this result confirmed that our proposed method can simultaneously reduce the effects of reverberation and additive noise.

B. ASR experiments on noisy reverberant speech

1) Feature extraction: The effectiveness of the proposed method was tested for simultaneous noise reduction and dereverberation as a front-end processor for ASR in noisy reverberant environments. The AURORA-2J database was used as speech material [9].We used 8,840 clean speech utterances to train the acoustic models. In addition, 100 clean speech sentences were used to produce noisy reverberant speech to simulate the noisy reverberant environments (convolution with RIRs and Gaussian white noise addition).

The restored sub-band temporal power envelope was obtained, and then the features were processed. The feature extraction includes a frame integration and log compression after the sub-band temporal power envelope extraction. Low-pass filtering with a forgotten parameter λ was used to smooth the envelope dips:

$$\overline{e}_{x,k}[p] = \lambda\overline{e}_{x,k}[p-1]\times(1-\lambda)\hat{e}_{x,k}[p], \quad (18)$$

where $\hat{e}_{x,k}[p]$ is the original restored sub-band power envelope, and $\overline{e}_{x,k}[p]$ is the smoothed output (k is the sub-band index, and p is the time frame index). In this study, λ was set to 0.99. In frame integration, a 32 ms frame length with a Hamming window and a 16 ms frame rate were used. After the integrated spectrum was obtained, log compression was carried out followed by the discrete cosine transform (DCT) for dimensional decorrelation. The first 12 dimension coefficients of the decorrelated log power spectrum were

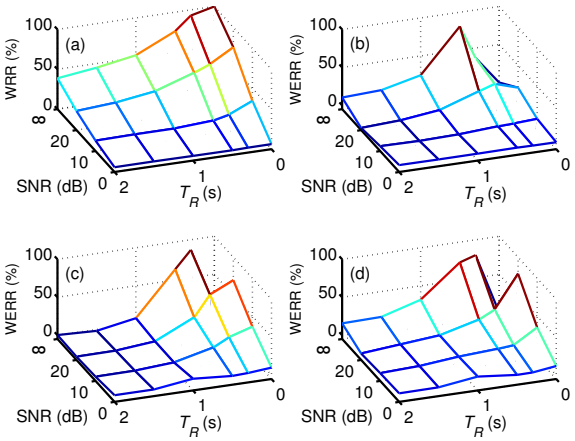

Fig. 4. Comparative evaluations of noisy and/or reverberant speech: (a) word recognition rates (WRR) using feature MFCC as baseline, (b) word error reduction rate (WERR) using feature CBFB, (c) WERR using feature CBFB_RASTA, and (d) WERR using feature CBFB_IMTF.

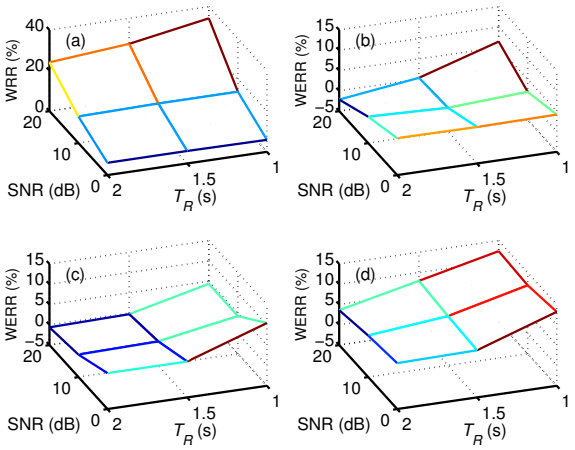

Fig. 5. Close-up of Fig. 4 as high noisy reverberant environments

used. By combining the log power energy, we obtained 13 dimensional static feature sets. Together with their first and second order delta dynamic values, 39 dimensional feature vectors were formed. HTK was used for training the HMM acoustic models, which were configured in the same way as in the AURORA-2J experiments [9].

2) Recognition results: The features extracted based on the proposed processing (Fig. 2), denoted as CBFB_IMTF, were tested for ASR. For comparison, the conventional Mel frequency cepstral coefficient (MFCC) feature was also tested under the same conditions; it served as the baseline. In addition, for a better understanding of the contribution of each stage in the proposed processing (Fig. 2), the sub-band (constant bandwidth filter) power envelope based cepstral feature (denoted as CBFB) [7] and the RASTA filtering on the CBFB feature (denoted as CBFB_RASTA), were also tested. The results are shown in Fig. 4 as MFCC is shown in the word recognition rate (WRR) and the another features are shown in the word error reduction rate (WERR) in noisy and/or reverberant environments. From this figure, we can see that CBFB, CBFB_RASTA, and CBFB_IMTF improved the recognition performance compared with the baseline (MFCC). Figure 5 showed close-up results of Fig. 4 to clarify the effectiveness of CBFB_IMTF in high noisy reverberant environments. From Figure 5, we can see that WRR of MFCC is considerable degraded, the other thing, the IMTF performed better than the RASTA filtering, while the RASTA filtering improved the CBFB. This figure shows proposed method is robust in high noisy reverberant environments. From Figure 5(d), WERR of $T_R = 1.0$ s are high level compared with the other feature that mean dereverberation is work well. Quantitatively, compared with the baseline, 6.28, 11.24, and 12.19 % relative improvements in the WERR were obtained for CBFB, CBFB_RASTA, and CBFB_IMTF, respectively (average of all tested noisy reverberant conditions).

V. Conclusion

We proposed a unified noise reduction and dereverberation framework based on the concept of MTF [8] in order to reduce background noise and reverberation. We then systematically evaluated the proposed method to restore the power envelope in a noisy and/or reverberant environment. Our results showed that the proposed method could reasonably restore the power envelope from noisy reverberant speech, based on the SNR and correlation improvement criteria described in section 4.1. When the proposed method was applied as a front-end for ASR systems in noisy and/or reverberant environments, it obtained a relative improvement of 12.19 % compared with a baseline performance in the error reduction rate (average of all tested noisy reverberant environments). In the future, we will evaluate the proposed method as a front-end for ASR systems in real noisy reverberant environments.

Acknowledgement

This work was partially supported by the Research Foundation for the Electrotechnology of Chubu (REFEC).

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] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 166-169, 1979.
[3] H. Hermansky, N. Morgan and H. G. Hirsch, " Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *ICASSP'93*, pp. 83-36, Minneapolis, 1993.
[4] K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, "Suppression of late reverberation effect on speech signal using log-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 854-857, 2009.
[5] 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.
[6] M. Unoki, K. Sakata, M. Furukawa 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.
[7] 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.*, vol. 29, no. 6, pp. 351-361, 2008.
[8] M. Unoki, Y. Yamasaki and M. Akagi, "MTF-based power envelope restoration in noisy reverberant environments," *Proc. EUSIPCO2009*, pp. 228-232, Glasgow, 2009.
[9] http://www.slp.cs.tut.ac.jp/CENSREC/, AURORA-2J database.