Acoust. Sci. & Tech. **29**, 6 (2008)

PAPER

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

Xugang Lu*, Masashi Unoki† and Masato Akagi‡

School of Information Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi, 923–1292 Japan

(*Received 5 September 2007, Accepted for publication 2 May 2008*)

Abstract: To reduce speech degradation in reverberant environments, we previously proposed a modulation-transfer-function (MTF)-based method of speech dereverberation. By considering the temporal modulation properties of speech, and the exponential decay properties of the power envelope of the impulse response of room acoustics, we obtained the following MTF relation: the sub-band power envelope of reverberant speech that can be represented as a convolution between the sub-band power envelope of clean speech and the power envelope of the impulse response of room acoustics. On the basis of the MTF relation, inverse MTF filtering can be applied to restoring the power envelopes of reverberant speech. Therefore, the impulse response of the room acoustics in this restoration dose not need to be measured at any time since we model the power envelope of the impulse response as an exponential decay function. We have tested how effective this method is as a front-end for automatic speech recognition (ASR) systems in artificial and real reverberant environments. Reverberant speech signals were created by simply convoluting clean speech (AURORA-2J database) with the artificially produced or real impulse responses of room acoustics. A method based on the auditory power spectrum was used as a baseline for comparison. Compared with the baseline, the proposed method for artificial reverberant environments produced a 35.67% relative improvement in the error reduction rate (on average, for reverberation times from 0.2 to 2.0 s), and for real reverberant environments (43 reverberant impulse responses), it produced a 25.78% relative improvement in the error reduction rate. The results demonstrate that our new approach can improve the robustness of speech-recognition systems in reverberant environments, and it performs better than conventional methods.

Keywords: Power envelope restoration, Speech recognition, Modulation transfer function, Power envelope inverse filtering

PACS number: 43.72.Dv, 43.60.Cg [doi:10.1250/ast.29.351]

1. INTRODUCTION

It is well known that reverberation smears significant features of speech so that its quality and intelligibility are degraded during communication. Restoring the original clean speech from the observed reverberant speech is, therefore, an important issue in various kinds of real speech-signal-processing applications, e.g., speech enhancement, hearing improvement, and automatic speech recognition (ASR). The ultimate goal of our work is to construct a blind method of speech dereverberation that can restore speech signals from reverberant speech without having to measure the impulse response of room acoustics, and that causes less loss due to the reverberation in the speech intelligibility and recognition rate.

Traditional methods such as spectral subtraction, Wiener filtering, and Bayesian estimation have been widely used [1–3] to improve speech quality and intelligibility when there is additive noise. These make use of different statistical properties of speech and noise to reduce noise components and to enhance the speech itself. Reverberation can generally be regarded as the convolution processing of acoustic speech and room acoustics. The temporal and spectral structures of speech in a reverberant environment are distorted by stochastic reverberation caused by the room-reflection characteristics, of walls, floors, and ceilings. It is difficult to distinguish clean-speech signals in a reverberant environment by using the statistical

*e-mail: xugang@jaist.ac.jp
†e-mail: unoki@jaist.ac.jp
‡e-mail: akagi@jaist.ac.jp

properties of the original speech and of the reflected speech because the speech is degraded by propagating alone multiple paths. Thus, traditional methods of reducing noise do not work well in reverberant environments.

Several algorithms for reducing convolution distortion have been proposed [4–7]. The two most well known are cepstral mean normalization (CMN) [8] and relative spectral (RASTA) filtering [9]. These can effectively reduce the distortions caused by short-term convolution channels, e.g., microphones, and telephone-transmission channels. In actual room acoustics, the reverberation time is far longer, and the properties of reverberant environments are that they are both time and spatially variant.

Several dereverberation algorithms using single or multiple microphones have been proposed for solving the room-reverberation problem [7,10–12]. In these approaches, the basic principle of dereverberation is to measure the impulse response of room acoustics or propagation channels, and then use inverse filtering to obtain dereverberated speech [10]. However, these methods require the impulse response of room acoustics for each dereverberation process to be remeasured if the conditions for room acoustics change.

Blind dereverberation, which does not need the impulse response of room acoustics to be measured, is preferred for real applications. One possible way of utilizing blind dereverberation is to use speech characteristics. For example, the harmonic structure of speech can be used [13,14]. This method needs the fundamental frequency from reverberant speech to be accurately estimated, which is difficult [15], and it does not seem to restore the consonant (nonharmonic) parts in speech.

In this study, we utilized the characteristics of speech and the impulse response functions of reverberant environments for speech dereverberation. Speech signals are highly temporally modulated in amplitude, and most of their intelligibility information is encoded in the temporal modulation envelope of all frequency bands [16]. This means that we need to restore the temporal modulated envelope of clean speech from reverberant speech to restore clean speech for recognition.

In the concept underlying the modulation transfer function (MTF), the impulse response of room acoustics is assumed to be a random variable with properties of exponential-decay temporally modulated and white-noise carriers [17,18]. In addition, in MTF-based speech dereverberation [7,19–21], the same assumption as for room acoustics is used and the speech signal is assumed to be a random variable with properties of temporally modulated and white noise carriers in each frequency band. On the basis of the results of stochastical analysis of the signals, the sub-band power envelope of reverberant speech can be exactly represented as the convolution between the sub-band power envelope of clean speech and the power envelope of the impulse response of room acoustics. To obtain sub-band power envelopes of clean speech, only inverse MTF (IMTF) filtering is needed because of the relationships between the power envelopes of sub-band reverberant speech, of sub-band clean speech, and of the impulse response of room acoustics. Therefore, this method of restoration does not need the impulse response of room acoustics to be measured to derive inverse filtering [21,22].

We previously proposed a sub-band power envelope inverse filtering method based on the MTF concept [21–23]. We tested its effectiveness in restoring the temporal power envelopes of reverberant signals using correlation and SNR measurements [21,22]. These tests demonstrated that the proposed method improves the accuracy of power-envelope restoration and improves speech intelligibility [23]. We also conducted a preliminary test on its capability to act as a front-end processor for speech recognition in artificial reverberant environments [24], and we found that it was extremely effective. However, we have not yet tested its effectiveness as a front-end processor for speech recognition in real reverberant environments. We evaluated how well the proposed method performed in real reverberant environments, including multipurpose halls, classic concert halls, lecture rooms, churches, event halls, and speech halls. In addition, we compared the new approach with some traditional front-end processes, including auditory-filter band processing, as well as CMN and RASTA processing.

The paper is organized as follows. Chapter 2 describes the underlying concept and the model for restoring the MTF-based sub-band power envelope. Chapter 3 describes how features are extracted for speech recognition. Chapter 4 describes the recognition experiments we conducted in artificial reverberant environments, and Chapter 5 describes the ones we undertook in real reverberant environments. In Chapter 6, we summarize the key points, discuss the improvements that are needed, and briefly describe future work.

2. CONCEPT, MODEL, AND ALGORITHM FOR RESTORING MTF-BASED SUB_BAND POWER ENVELOPE

2.1. Concept and Model

The MTF concept was proposed by Houtgast and Steeneken [17] to account for the relationship between the transfer function in an enclosure in terms of input and output signal envelopes and the characteristics of the enclosure such as reverberation. This concept was introduced as a measure in room acoustics for assessing the effect of an enclosure on speech intelligibility [17]. The complex MTF is defined [18] as

$$M(\omega) = \frac{\int_0^\infty h^2(t)\exp(-j\omega t)dt}{\int_0^\infty h^2(t)dt}, \quad (1)$$

where $h(t)$ is the impulse response of the room and ω is the radian frequency. For room acoustics, a well-known stochastic approximation of the impulse response is defined [18] as

$$h(t) = e_{\mathrm{h}}(t)\boldsymbol{n}_1(t) = a\exp(-6.9t/T_{\mathrm{R}})\boldsymbol{n}_1(t), \quad (2)$$

where $e_{\mathrm{h}}(t)$ is the exponential decay temporal envelope, a is a constant amplitude, T_{R} is the reverberant time defined as the time required for the power of $h(t)$ to decay by 60 dB, and $\boldsymbol{n}_1(t)$ is white noise as a random variable (uncorrelated carrier) [21].

The corresponding MTF is obtained using

$$|M(\omega)| = \left[1 + \left(\omega\frac{T_{\mathrm{R}}}{13.8}\right)^2\right]^{-1/2}. \quad (3)$$

For modulation frequency ω of the temporal envelope, Eq. (3) can be regarded as the modulation index, i.e., the degree of relative fluctuation in the normalized amplitude with respect to the modulation frequency. On the basis of this characteristic, T_{R} can be predicted from a specific modulation frequency by using the MTF.

We modeled what effect room acoustics has on speech signals on the basis of the MTF concept. The convolution distortion in each sub-band is written as

$$y_n(t) = x_n(t) * h(t), \quad n = 1, 2, \cdots, N, \quad (4)$$

where $y_n(t)$ and $x_n(t)$ correspond to the reverberant and clean speech signals in the sub-band, n is the sub-band index, and N is the total number of sub-bands. Using the temporal modulation properties of the speech signal, we model the sub-band speech, $x_n(t)$, as

$$x_n(t) = e_{x,n}(t)\boldsymbol{n}_2(t). \quad (5)$$

The temporal envelope of sub-band n is $e_{x,n}(t)$. In Eqs. (2) and (5), $\boldsymbol{n}_1(t)$ and $\boldsymbol{n}_2(t)$ are mutually independent random variables that satisfy

$$\langle \boldsymbol{n}_k(t)\boldsymbol{n}_k(t-\tau)\rangle = \delta(\tau), \quad k = 1, 2, \quad (6)$$

where $\langle\cdot\rangle$ is the ensemble average operator. Using Eqs. (4)–(6), we can calculate the power envelope of $y_n(t)$ as

$$\langle y_n^2(t)\rangle = e_{y,n}^2(t) = e_{x,n}^2(t) * e_{\mathrm{h}}^2(t) \quad (7)$$

(for details, see Appendix in [21]). This equation shows that the restoration of $e_{x,n}^2(t)$ can be completed by deconvoluting $e_{y,n}^2(t)$ with $e_{\mathrm{h}}^2(t)$. To cope with these signals in computer simulation, the variables are transformed from a continuous signal into a discrete signal on the basis of sampling theorems, such as $e_{x,n}^2[m]$, $e_{\mathrm{h}}^2[m]$, $e_{y,n}^2[m]$, $x[m]$, $h[m]$, and $y[m]$ (m is the number of samples). The transfer functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to be the respective z-transforms of $e_x^2[m]$, $e_{\mathrm{h}}^2[m]$, and $e_y^2[m]$. The input-output relationship for deconvolution can be represented as

$$E_{x,n}(z) = \frac{E_{y,n}(z)}{E_h(z)} = \frac{E_{y,n}(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_{\mathrm{R},n}\cdot f_{\mathrm{s}}}\right)z^{-1}\right\}, \quad (8)$$

where f_{s} is the sampling frequency. The power envelope of sub-band signal $e_{x,n}^2(t)$ can be restored using the inverse z-transform of $E_{x,n}(z)$. In Eq. (8), we only need to estimate parameters $T_{\mathrm{R},n}$ and a. Here, the parameter of the inverse MTF filter related to reverberant time $T_{\mathrm{R},n}$ is assumed to be a function of n since it is dependent on the sub-band, and is independently estimated from each sub-band.

2.2. Algorithm

The algorithm for inverse filtering of the sub-band power envelope was developed on the basis of the analysis above. The processing scheme for inverse filtering of the sub-band power envelope is outlined in Fig. 1. In the processing scheme, observed signal $y(t)$ (a pre-emphasized signal of the original signal with a coefficient of 0.97) is decomposed into a series of frequency sub-bands; envelope detectors then extract temporal modulation envelopes $e_{y,n}^2(t)$. Considering the co-modulation characteristics of speech signals in the sub-bands [22], we deliberately designed a series of FIR-type band-pass filters with a constant bandwidth (100 Hz was chosen in this study) for the decomposition (see Subchapter 4.1 for bandwidth selection). Thus, this filterbank is referred to as a constant-bandwidth filterbank (CBFB) in this paper. The extracted envelopes are used for inverse filtering of the power envelope, which is controlled using the estimated parameters of $T_{\mathrm{R},n}$ and a (referred to as $\hat{T}_{\mathrm{R},n}$ and $\hat{a}$). The final output is the restored or dereverberated power envelope, $\hat{e}_{x,n}^2(t)$, for all sub-bands. The implementation is detailed in the three steps in Subsections 2.2.1, 2.2.2, and 2.2.3.

2.2.1. Power envelope extraction

The power envelopes in the sub-bands are extracted by low-pass filtering of the Hilbert transform of the sub-band signals [21,22].

$$\hat{e}_{y,n}^2(t) = \mathbf{LPF}[|y_n(t) + j\mathbf{Hilbert}(y_n(t))|^2] \quad (9)$$

Here $\mathbf{LPF}[\cdot]$ is a low-pass filtering operator and $\mathbf{Hilbert}(\cdot)$ is the Hilbert transform. We set the cut-off frequency of the low-pass filtering to 20 Hz to retain most of the important modulation information for speech perception [22].

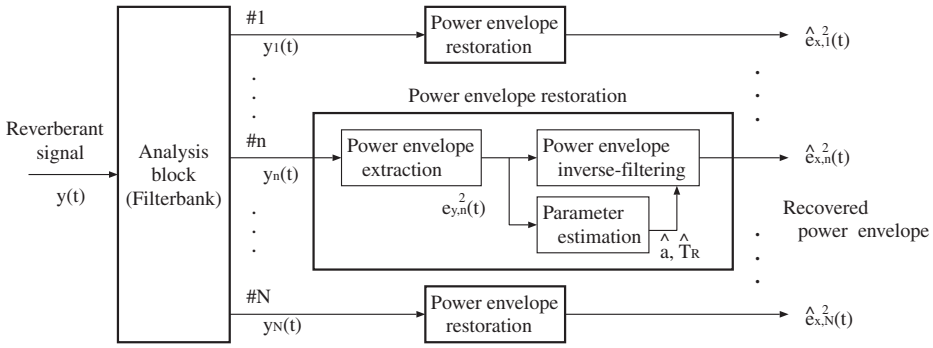

Fig. 1 Sub-band power envelope method of inverse filtering.

2.2.2. Parameter estimation

The $T_{\mathrm{R},n}$ and a in Eq. (8) are estimated using Unoki *et al.*'s formulas [21]:

$$\hat{T}_{\mathrm{R},n} = \max\left(\arg\min_{T_{\mathrm{R},n}} \int_0^T \left| \min(\hat{e}^2_{x,n,T_{\mathrm{R},n}}(t), 0) \right| dt \right), \qquad (10)$$

and

$$\hat{a} = \sqrt{1 \Big/ \int_0^T \exp\left(-\frac{13.8t}{\hat{T}_{\mathrm{R},n}} \right) dt}, \qquad (11)$$

where T is the signal duration, and $\hat{e}^2_{x,n,T_{\mathrm{R},n}}(t)$ represents the candidates of the restored power envelope as a function of $T_{\mathrm{R},n}$. The reverberant time is constrained as $T_{\mathrm{R},n,\min} < T_{\mathrm{R},n} < T_{\mathrm{R},n,\max}$. $T_{\mathrm{R},n,\min}$ and $T_{\mathrm{R},n,\max}$ are the lower and upper bounds of $T_{\mathrm{R},n}$ (the former was set to 0 s and the latter to 3 s in our study). Estimating the reverberant time using Eq. (10) means finding the maximum argument of $T_{\mathrm{R},n}$ from a time point to obtain the minimum area of the estimated inverse power envelope with a constraint of no less than zero. Equations (10) and (11) are described in detail elsewhere [21,22].

2.2.3. Power envelope inverse filtering

After the power envelopes ($e^2_{y,n}(t)$) and the parameters of room acoustics ($\hat{T}_{\mathrm{R},n}$ and $\hat{a}$) are obtained, the power envelopes are inverse filtered using Eq. (8) to restore the power envelopes of dereverberated speech in the sub-bands ($e^2_{x,n}(t)$). Here, the restored power envelope of the dereverberated speech in a sub-band is denoted as $\hat{e}^2_{x,n}(t)$.

Figure 2 shows an example of the effect of restoring the sub-band power envelope on reverberant speech. The stimulus was a Japanese sentence (/aikawarazu/) uttered by a male speaker (panel (a)), and reverberant speech occurred when $T_{\mathrm{R}} = 1.0$ s (panel (b)). The power envelopes of only odd-numbered channels are plotted in this figure (channels #1, #3, #5, ..., #39). All pairs of the power envelopes (solid and dashed lines) have also been plotted by normalizing them to qualitatively compare matches between the power envelopes of the original and the restored envelopes. Comparing the sub-band envelopes in panels (c) and (d) in Fig. 2, we can see that, based on the processing procedures, the sub-band power envelope of reverberant speech can be restored to be close to that of clean speech.

We have already investigated the restoration accuracy using two methods. The first was the correlation between the restored sub-band power envelopes of reverberant speech and the sub-band power envelope of original clean speech, while the second was the SNR, where S is the original sub-band power envelope, and N is the difference between the original and the restored sub-band power envelope (see Eqs. (14) and (15) in [21,22]). Many improvements have been reported for both evaluations (see Fig. 10 in [22] for details). At the same time, we found that over- and under estimating the reverberant time does not optimally restore the temporal power envelope based on the MTF concept (see Appendix I for a further illustration of the effect of over- and under estimating of $T_{\mathrm{R},n}$).

3. FEATURE EXTRACTION FOR SPEECH RECOGNITION

We tested the effectiveness of the proposed algorithm for dereverberation as a front-end processor for ASR of reverberant speech. We used clean speech from the AURORA-2J database as the speech material [25], and 8,840 clean speech sentences to train the acoustic models. We used 1,001 clean speech sentences to produce reverberant speech to test recognition in reverberant environments by convolving the speech signals with the impulse responses of room acoustics. As the sampling frequency, f_{s}, was 8 kHz, we used 40 sub-band channels ($N = 40$) to cover the frequency region from 0 to 4 kHz. After the restored power envelopes had been obtained from the processing blocks represented in Fig. 1, the speech feature was further extracted as illustrated in Fig. 3.

In Fig. 3, the first block is for smoothing which comprises frame integration and log compression. Because the inverse filtering of power envelopes is a high-pass process, low-pass filtering with a forgotten parameter, λ (in the region between 0 and 1), was used to smooth the envelope dips in each sub-band:

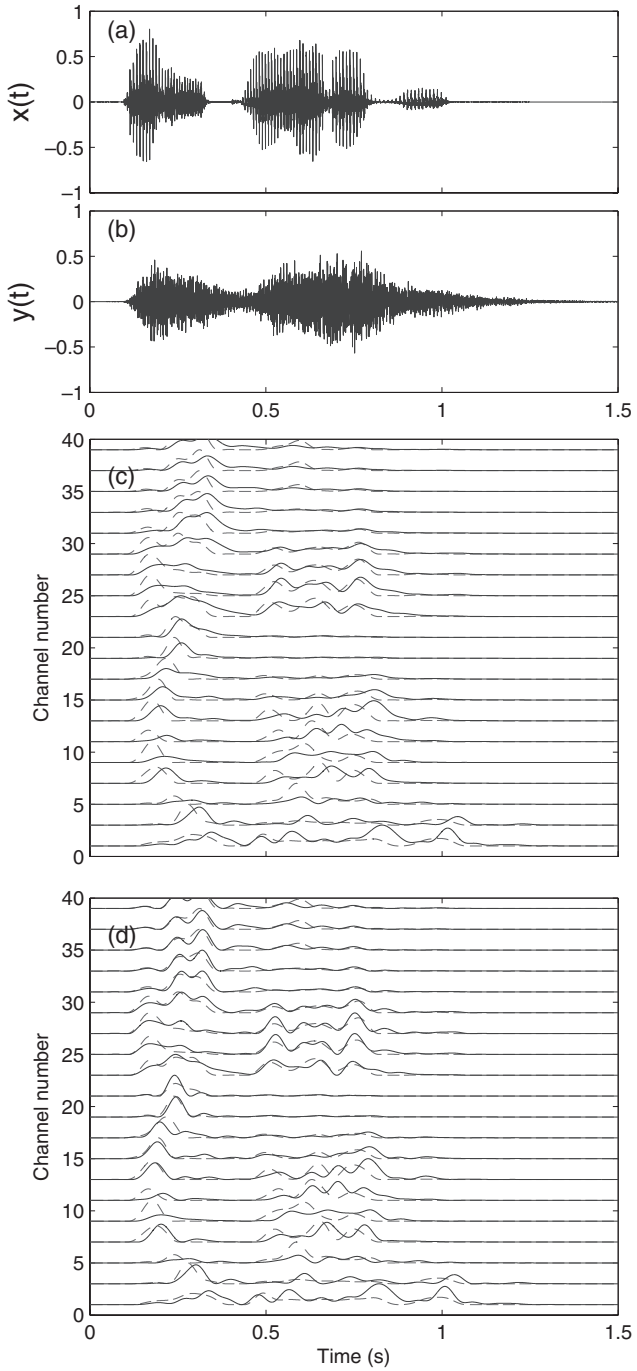

Fig. 2 Effect of restoring temporal envelope by inverse MTF filtering: (a) clean speech, (b) reverberant speech with $T_R = 1.0$ s, (c) no dereverberation processing (solid lines), and (d) restored envelope using inverse filtering (solid lines). The dotted lines correspond to the power envelopes of the original clean speech.

$$\bar{e}_{x,n}[m] = \lambda \bar{e}_{x,n}[m-1] + (1-\lambda)\hat{e}_{x,n}[m], \qquad (12)$$

where $\hat{e}_{x,n}[m]$ is the original restored sub-band power envelope, and $\bar{e}_{x,n}[m]$ is the smoothed output. In this study, we set λ to 0.98. To integrate the frames, we used a 32 ms frame length with a Hamming window and a frame rate of 16 ms. After the integrated spectrum was obtained, log compression was carried out. The discrete cosine transform (DCT) was used for dimensional decorrelation. The first 12 dimensions of the decorrelated log power spectrum were used (the zeroth-order coefficient was discarded). Combining the log power energies, we obtained 13-dimensional static feature sets. Together with their first- and second-order delta dynamic values, 39-dimensional feature vectors were formed. The acoustic models consist of ten digits, one silence and short-pause HMM models which are the same as those used in the AURORA-2J experiments [25]. Each digit model had 18 states with 16 output distributions. The silence model had five states with three distributions, and the short-pause model had three states with one distribution. Each distribution of digits had 20 Gaussian mixtures, while those of the silence and short-pause models had 36 Gaussian mixtures. The HTK speech toolkit [26] was used for training the HMM acoustic models.

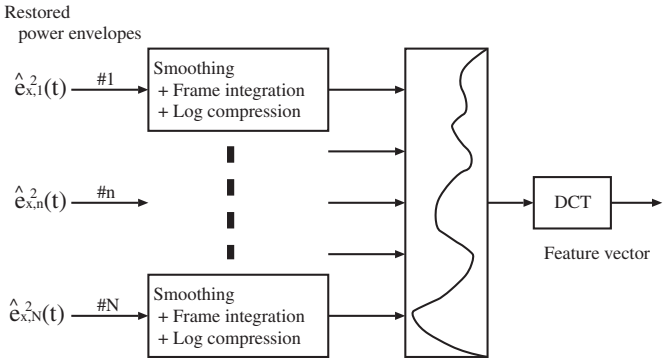

Fig. 3 Extraction of speech features based on restored power envelope in all sub-bands.

4. RECOGNITION EXPERIMENTS IN ARTIFICIAL REVERBERANT ENVIRONMENTS

As previously mentioned, we tested the recognition of our proposed method in artificial reverberant environments by using 1,001 clean speech sentences to produce reverberant speech. The speech signals were convolved with the artificial impulse responses of room acoustics (produced using Eq. (2)) with a reverberation time of 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 or 2.0 s. In total, we used 1,001 clean speech signals and $1{,}001 \times 10$ reverberant speech signals.

For comparison, we also tested the performance of a conventional method of feature extraction based on a computational auditory model under the same conditions. There are two steps in processing by conventional feature extraction: extraction of the sub-band power envelope using the auditory filterbank illustrated in Fig. 4, and post-processing of speech feature extraction on the basis of the sub-band power envelopes using the same process as illustrated in Fig. 3.

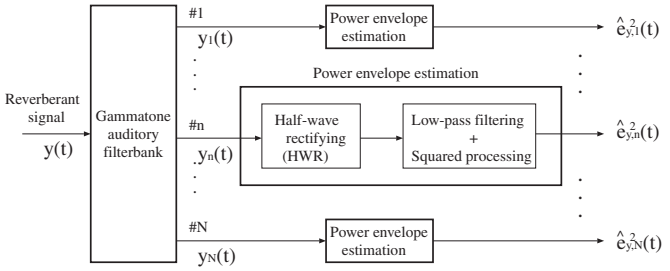

Fig. 4 Extraction of sub-band power envelope based on auditory filterbank.

As seen in Fig. 4, the speech signal is first decomposed using a gammatone auditory filterbank with an equivalent rectangular bandwidth (ERB). Half-wave rectification and low-pass filtering are then used to extract the sub-band temporal amplitude envelopes. The power envelopes are obtained by square processing of the amplitude envelopes. The filterbank can be regarded as a constant-Q filterbank (CQFB). On the basis of the sub-band power envelopes, we obtain the auditory cepstral feature vector, which is denoted as CQFB, using the processing block in Fig. 3. In our preliminary experiments on speech recognition, we find that CQFB exhibits equivalent performance to or slightly better performance than the Mel Frequency Cepstral Coefficient (MFCC) representation. Considering that our new approach involves sub-band filtering and extraction processing of the temporal envelope, we choose the performance of CQFB as a baseline for comparison. Two conventional post-processing methods, i.e., RASTA filtering [9] and CMN [8], are also implemented in this study to deal with convolution distortion. In our study, the two processes are used on utterance-based cepstral temporal trajectories. Consequently, the features extracted are denoted here as Fea_RASTA and Fea_CMN (where "Fea" is either CQFB or CBFB).

4.1. Effect of Bandwidth of Band-Pass Filtering

Constant-Q band-pass-filtering (ERB or Mel)-based extraction of features is widely accepted as being more robust under most additive noise conditions than constant band-pass filtering. However, because our MTF-based dereverberation is based on the inverse process of the sub-band power envelope, this envelope should have co-modulation characteristics in the sub-bands while satisfying the MTF concept. We carefully choose the bandwidth by considering the trade-off in their characteristics [22,23]. In an earlier experiment, we tested what the effect of using ERBs of gammatone auditory filters (i.e., constant-Q) and constant filter bandwidths would have on recognition [24]. We found that a constant bandwidth of 100 Hz is more suitable for satisfying the envelope co-modulation properties and the MTF concept. This result is consistent with that

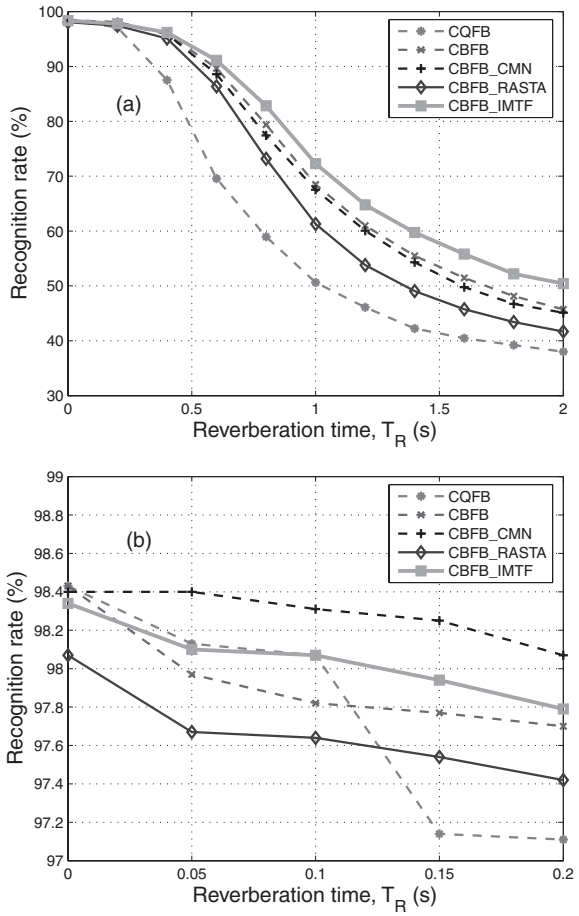

Fig. 5 Comparative evaluations of reverberant speech recognition rates: (a) whole evaluation and (b) close-up of plot in range from 0.0 to 0.2 s.

reported elsewhere [19,20]. Therefore, in our MTF-based dereverberation experiments, we used band-pass filters with a 100 Hz bandwidth. In addition, we compared the results of recognition using CQFB and CBFB, and found that CBFB outperformed CQFB [27].

4.2. Comparison with Conventional Feature Extraction Methods

We simulated speech recognition using many types of feature vectors, i.e., CQFB, CBFB, CBFB_CMN, CBFB_RASTA, and CBFB_IMTF.

The recognition for short reverberation times ($T_R <$ 0.2 s) was best, as can be seen from the magnified plot in Fig. 5(b). As seen in the figure, the speech-recognition rate decreased as the reverberation time increased; the rate of decrease was particularly high when the reverberation time was long ($T_R > 0.2$ s). When it was short ($T_R < 0.2$ s), all the features performed well (recognition rate $> 97\%$). The CBFB-based feature performed better for long reverberant times, $T_R > 0.15$ s, and slightly worse for short reverberant times, $T_R < 0.15$ s, than the CQFB-based feature. Also, as shown in Fig. 5, CBFB_RASTA per-

formed worse than CBFB alone. CBFB_CMN performed slightly worse or almost the same as CBFB alone (except for short reverberant times, $T_R < 0.2$ s). However, CBFB_IMTF consistently improved the performance of CBFB alone. We also tested how well CQFB_CMN, CQFB_RASTA, and CQFB_IMTF performed. They performed worse than CBFB alone (see Appendix II for the results). Consequently, we did not use CQFB_CMN, CQFB_RASTA, or CQFB_IMTF for comparison in later experiments. In addition, we found that adding CMN or RASTA processing to either CQFB or CBFB did not improve the recognition rate in our experiments, and sometimes even decreased the performance. Therefore, in this study, the performance of CQFB was used as a baseline. A relative improvement (RI) in performance was adopted as the measure of the improvement in recognition in different reverberant environments from different baselines, and was defined [25] as

$$\mathrm{RI} = \frac{(\mathrm{TRR} - \mathrm{BRR})}{(1 - \mathrm{BRR})} \times 100 \quad (\%), \tag{13}$$

where "TRR" and "BRR" denote the testing recognition rate and baseline recognition rate. With this definition, the proposed CBFB and CBFB_IMTF yielded relative improvements of 28.64% and 35.67% on average (for $0.2\,\mathrm{s} < T_R < 2\,\mathrm{s}$), respectively, in the error reduction rate compared with CQFB.

5. RECOGNITION EXPERIMENTS IN REAL REVERBERANT ENVIRONMENTS

We then tested our method on speech recognition in many real reverberant environments (43 halls, rooms, and theaters [28]) under various conditions. The reverberant speech signals were obtained from the convolutions between clean speech signals and the impulse responses of the environments (the sampling rate of the impulse responses of the environments were sub-sampled to 8 kHz to adjust it to the sampling rate of the speech database). The speech corpus, features, and acoustic models were the same as those used for the artificial reverberant environments described in Chapter 4. The characteristics of the reverberant environments and the speech recognition rates are listed in Table 1.

Rooms and halls constructed of different materials and with different configurations have vastly different reverberant characteristics. As listed in Table 1, the reverberant times for rooms and halls in which we conducted our tests ranged widely, from 0.36 to 3.62 s. The middle-six columns list speech-recognition rates with the highest rates for each room or hall marked in bold. The rightmost column indicates the relative improvements in the error reduction rate of the CBFB_IMTF feature compared with that of the CQFB feature. The CQFB feature performs almost the same as or slightly better than MFCC (on average).

The CBFB-based features outperformed the CQFB-based features. CBFB_RASTA and CBFB_CMN showed no improvements in performance compared with that of CBFB alone. The CBFB_IMTF-based feature had the highest recognition in almost every case. On average, CBFB and CBFB_IMTF had relative improvements of 15.74% and 25.78% compared with CQFB. The table also indicates how the differences in the acoustic characteristics of the various environments affected the speech-recognition rate. There was almost no degradation in speech recognition for the meeting room, wooden house, living room, or movie theater (recognition rate $> 90\%$) because these environments had originally been designed to minimize reverberant properties.

In contrast, there was significant degradation in speech recognition in the classic concert hall environments, probably because these had been designed to emphasize reverberant properties to enhance the audience's musical enjoyment. In Table 1, we can see the effects of reverberant configurations with various reflective properties of room acoustics, e.g., with or without reflective boards, absorptive boards, and absorptive curtains. In addition, we can also see the effects of different spatial recording locations and of the distance between the microphone and sound sources had. A more detailed investigation of these effects is beyond the focus of this study.

6. DISCUSSION AND CONCLUSION

Our analysis and experiments demonstrated that our MTF-based sub-band power envelope extraction and inverse filtering algorithm improves the robustness of speech recognition for reverberant speech. The results revealed that (1) constant-Q band-pass processing or MFCC presented no advantages in improving ASR in reverberant environments. As Shannon and Paliwal [29] pointed out, auditory motivated band-pass processing does not have any advantages over other types of band-pass processing. The band-pass filters in our study were deliberately designed to fulfill the co-modulation properties of speech. The results also revealed that (2) considering the exponential decay properties of an impulse response in a reverberant environment and the temporal modulation properties of speech, we can estimate the sub-band temporal power envelope of speech to some degree without having to measure the impulse response of room acoustics, thereby improving the ASR of reverberant speech, and that (3) in real reverberant environments, the proposed estimates of the sub-band temporal envelope with inverse filtering based on dereverberation consistently improves ASR (25.78% relative improvement on average).

Upon comparing the recognition rates of CBFB_IMTF and CBFB in Table 1, we find that adding inverse filtering

Table 1 Reverberant speech recognition rates (%) in actual reverberant environments. IRdata No. indicates File No. in SMILE2004 [28]. The reverberation time, T_R, was determined as the average of all T_Rs on the transfer function at 125 Hz to 8 kHz in octave frequencies. "RI_CQFB" and "RI_CBFB" mean the relative improvement in the error reduction rate of the CBFB_IMTF feature compared with those of CQFB and CBFB features. MPH: Multi-purpose hall; CCH: Classic concert hall; GSH: General speech hall, RB: Reflective board, AB: Absorptive board, AC: Absorptive curtain.

Room condition (Impulse response)	IRdata No.	T_R (s)	MFCC	CQFB	CBFB	CBFB_ CMN	CBFB_ RASTA	CBFB_ IMTF	RI_ CQFB	RI_ CBFB
MPH 1 (with RB)(capacity: 2,000 m^3)	301	1.09	42.55	45.56	52.44	57.63	48.51	**60.30**	*27.08*	*16.53*
MPH 1 (without RB)	302	0.80	55.39	54.31	68.52	71.85	66.17	**74.33**	*43.82*	*18.46*
MPH 2 (with RB)(capacity: 5,700 m^3)	303	1.44	32.88	36.60	40.62	39.70	32.64	**45.41**	*13.90*	*8.07*
MPH 2 (without RB)	304	1.04	39.70	43.51	47.56	45.49	36.20	**52.38**	*15.70*	*9.19*
MPH 3 (with RB)(capacity: 7,200 m^3)	305	1.93	30.70	33.40	33.80	35.31	31.26	**39.31**	*8.87*	*8.32*
MPH 3 (without RB)	306	1.35	42.12	43.48	46.52	53.42	47.50	**54.19**	*18.95*	*14.34*
MPH 4 (with AB)(capacity: 12,000 m^3)	307	1.42	55.70	55.07	69.63	74.24	71.05	**75.87**	*46.29*	*20.55*
MPH 4 (without AB)	308	1.54	52.44	53.42	67.02	71.08	66.78	**73.10**	*42.25*	*18.44*
MPH 5 (capacity: 14,000 m^3)	319	1.47	46.55	47.28	61.38	59.84	54.71	**64.04**	*31.79*	*6.89*
MPH 6 (capacity: 19,000 m^3)	321	2.16	40.13	42.83	49.95	49.43	47.99	**54.49**	*20.40*	*9.07*
CCH 1 (capacity: 5,600 m^3)	309	2.35	27.72	34.20	35.19	33.50	28.92	**35.92**	*2.61*	*1.13*
CCH 1 ($d = 6$ m)	310	2.34	30.09	35.65	39.88	37.03	33.22	**42.74**	*11.02*	*4.76*
CCH 1 ($d = 11$ m)	311	2.35	30.40	35.22	37.67	35.34	33.19	**43.17**	*12.27*	*8.82*
CCH 1 ($d = 15$ m)	312	2.39	30.58	35.37	39.73	38.44	35.55	**45.47**	*15.63*	*9.52*
CCH 1 ($d = 19$ m)	313	2.38	27.82	33.93	36.17	34.30	32.36	**40.56**	*10.03*	*6.88*
CCH 2 (capacity: 6,100 m^3)	314	1.14	40.34	44.34	50.60	58.12	49.59	**59.84**	*27.85*	*18.17*
CCH 3 (capacity: 20,000 m^3)	315	1.96	35.00	36.81	37.73	42.80	39.12	**46.33**	*15.07*	*13.81*
CCH 4 (with AC)(capacity: 7,100 m^3)	316	1.92	41.23	41.42	50.02	49.95	46.15	**54.38**	*22.12*	*8.72*
CCH 4 (without AC)	317	2.55	34.33	36.72	41.97	41.14	37.15	**44.43**	*12.18*	*4.24*
CCH 5 (capacity: 17,000 m^3)	323	2.32	31.78	37.70	38.29	34.85	32.58	**44.09**	*10.19*	*9.40*
CCH 6 (1F front) (capacity: 17,000 m^3)	324	1.77	37.73	41.42	43.57	42.55	38.38	**53.45**	*20.54*	*17.51*
CCH 6 (2F side)	325	1.74	40.13	44.18	47.87	46.27	42.25	**55.14**	*19.63*	*13.95*
CCH 6 (3F)	326	1.69	34.73	38.23	44.34	43.11	41.42	**52.69**	*23.41*	*15.00*
Lecture room (with flatter echo)	201	1.36	46.76	45.72	60.85	**70.31**	67.58	68.53	*42.02*	*19.62*
Theater hall (capacity: 3,900 m^3)	318	0.85	46.24	48.82	60.55	60.39	53.39	**63.68**	*29.03*	*7.93*
Meeting room (capacity: 130 m^3)	401	0.62	77.43	72.24	89.10	91.25	89.16	**91.62**	*69.81*	*25.87*
Lecture room (capacity: 400 m^3)	402	1.12	55.85	53.18	70.83	**81.12**	78.75	80.32	*57.97*	*32.53*
Lecture room (capacity: 2,400 m^3)	403	1.09	57.48	51.30	68.35	**83.97**	80.75	78.85	*56.57*	*33.18*
GSH (capacity: 11,000 m^3)	404	1.54	40.44	44.89	51.58	46.58	44.55	**54.34**	*17.15*	*5.70*
Church 1 (capacity: 1,200 m^3)	405	0.71	57.35	56.95	70.34	76.60	72.43	**77.56**	*47.87*	*24.34*
Church 2 (capacity: 3,200 m^3)	406	1.30	33.71	37.21	41.42	40.87	30.52	**42.49**	*8.41*	*1.83*
Event hall 1 (capacity: 28,000 m^3)	407	3.03	27.51	31.19	33.40	33.40	30.80	**36.87**	*8.25*	*5.21*
Event hall 2 (capacity: 41,000 m^3)	408	3.62	28.77	32.98	35.62	37.27	34.88	**41.63**	*12.91*	*9.34*
Gym 1 (capacity: 12,000 m^3)	409	2.82	21.61	26.59	29.08	27.88	25.39	**30.09**	*4.77*	*1.42*
Gym 2 (capacity: 29,000 m^3)	410	1.70	32.51	37.33	39.98	41.60	36.29	**48.23**	*17.39*	*13.90*
Living room (wooden)(capacity: 110 m^3)	411	0.36	89.81	86.40	**98.31**	96.75	95.30	96.90	*77.21*	*−83.93*
Movie theater (capacity: 560 m^3)	412	0.38	88.36	84.22	93.49	**95.95**	92.85	93.18	*56.78*	*−4.76*
Antrum (capacity: 4,000 m^3)	413	1.57	35.19	36.91	39.70	43.97	36.08	**48.60**	*18.53*	*14.76*
Tunnel (capacity: 5,900 m^3, length: 120 m)	414	2.72	28.52	25.05	25.33	26.76	**35.06**	33.87	*11.77*	*11.44*
Concourse in train station	415	1.95	36.66	39.64	44.06	**46.18**	34.48	45.93	*10.42*	*3.34*
GSH 2 (1F front)	416	1.53	38.26	41.45	48.33	46.88	42.80	**56.13**	*25.07*	*21.10*
GSH 2 (1F center)	417	1.49	34.26	37.67	45.13	44.98	41.26	**51.77**	*22.62*	*12.10*
GSH 2 (1F balcony)	418	1.40	39.73	39.05	54.41	59.81	56.19	**65.18**	*42.87*	*23.62*

to CBFB does not greatly improve the recognition rate (only a 5.77% absolute improvement in the recognition rate on average). The recognition rates are still low for many reverberant conditions. This suggests that we must reconsider how some things are handled from both model and implementation aspects. For example, speech is assumed to be a temporal envelope modulated with a Gaussian white-noise carrier signal. However, the carrier may be non-Gaussian for a real speech signal. Therefore, one direction to take to achieve improvement is to relax the model assumptions. In terms of implementation, dereverberation is accomplished using the estimated reverberation time in each sub-band independently. If there is even a small error in the estimates, the extracted feature may differ greatly from the actual ones owning to temporal misalignment between sub-bands. A more accurate method of estimating

the reverberation time is thus needed. We must also find a more accurate method of estimating the sub-band temporal power envelopes because the inverse MTF filtering for dereverberation is based on these envelopes. We need a way to estimate the sub-band temporal power envelope by stochastic signal processing for both Gaussian and non-Gaussian white-noise carriers. In our experiments, reverberant speech was obtained by manual convolution between the speech and the artificial or real impulse response of the room acoustics. However, we must consider real reverberant speech, which should be recorded in a reverberant environment. In addition, apart from convolution-distortion, additive noise in real reverberant environments may cause speech to degrade.

Finally, as we mentioned in Chapter 1, the impulse response of a reverberant environment may be time-variant. Our proposed inverse filtering, in Eq. (8), can be used as time-variant filtering if we can estimate the instantaneous reverberant time. In our current study, the reverberant time was estimated from each utterance. Some utterances in the speech database were long (about 4.2 s), and others were short (about 0.8 s), and, on average, they were about 2 s for the data corpus. It is possible to estimate the reverberant time using a short period of speech, such as in described [23]. In the future, we intend to deal with real recorded speech in reverberant environments by adapting and modifying our MTF-based processing model.

ACKNOWLEDGEMENTS

This work was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (Nos. 18680017 and 18700172). 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. We would like to thank ATR Spoken Language Translation Research Laboratories for permitting us to use the AURORA-2J data.

REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-33**, 443–445 (1985).

[3] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process.*, **10**, 1043–1051 (2003).

[4] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing* (Marcel Dekker, Inc., New York, 1991).

[5] T. Takiguchi, S. Nakamura and K. Shikano, "Hands-free speech recognition by HMM composition in noisy reverberant environments," *IEICE Trans. D-II*, **J79-D-II**, 2047–2053 (1996) (in Japanese with English figure captions).

[6] S. Nakagawa, "A survey on automatic speech recognition," *IEICE Trans. D-II*, **J83-D-II**, 433–457 (2000) (in Japanese with English figure captions).

[7] K. Kinoshita, T. Nakatani and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," *Proc. ICASSP 2006*, Vol. I, pp. 817–820 (2006).

[8] F. Liu, R. Stern, X. Huang and A. Acero, "Efficient cepstral normalization for robust speech recognition," *Proc. ARPA Human Language Technology Workshop* (1993).

[9] H. Hermansky, N. Morgan and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *ICASSP '93*, pp. 83–86 (1993).

[10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-36**, 145–152 (1988).

[11] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. (Springer-Verlag, Berlin, 2001).

[12] J. B. Allen, D. A. Berkley and J. Blauert, "Multi-microphone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, **62**, 912–915 (1977).

[13] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP '03*, Vol. 1, pp. 92–95 (2003).

[14] T. Nakatani, M. Miyoshi and K. Kinoshita, "Blind dereverberation of monaural speech signals based on harmonic structure," *IEICE Trans. D-II*, **J88-D-II**, 509–520 (2005) (in Japanese with English figure captions).

[15] M. Unoki, T. Hosorogiya and Y. Ishimoto, "Comparative evaluations of robust and accurate F0 estimates in reverberant environments," *Proc. ICASSP 2008*, pp. 4569–4572 (2008).

[16] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, **270**, 303–304 (1995).

[17] 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).

[18] M. R. Schroeder, "Modulation transfer function: Definition and measurement," *Acustica*, **49**, 179–182 (1981).

[19] S. Hirobayashi, H. Nomura, T. Koike and M. Tohyama, "Speech waveform recovery from a reverberant speech signal using inverse filtering of the power envelope transfer function," *IEICE Trans. A*, **J81-A**, 1323–1330 (1998) (in Japanese with English figure captions).

[20] S. Hirobayashi and T. Yamabuchi, "Validation of blind dereverberation using power envelope inverse filtering and filter banks," *IEICE Trans. A*, **J83-A**, 1029–1033 (2000) (in Japanese with English figure captions).

[21] 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**, 232–242 (2004).

[22] 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.*, **25**, 243–254 (2004).

[23] M. Unoki, M. Toi and M. Akagi, "Development of the MTF based speech dereverberation method using adaptive time-frequency division," *Proc. Forum Acusticum 2007*, pp. 51–56 (2005).

[24] X. Lu, M. Unoki and M. Akagi, "A robust feature extraction based on the MTF concept for speech recognition in

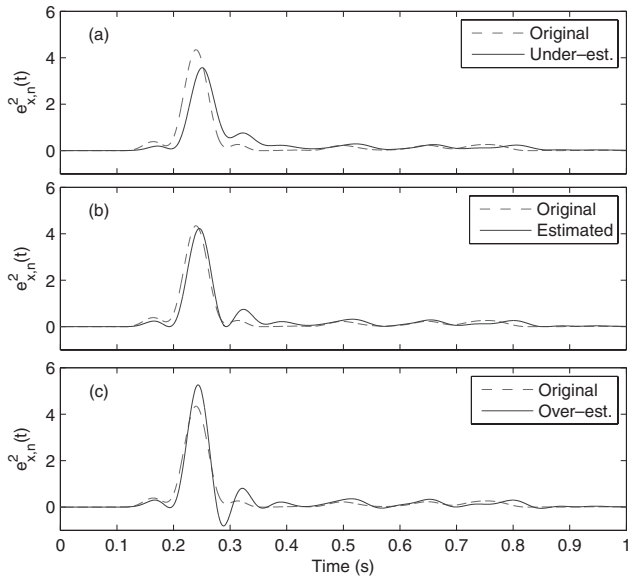

Fig. 6 Effects of over- and under estimation of reverberant time T_R on restored power envelope based on (a) under estimated reverberant time (solid line), (b) reverberant time estimated using proposed method (solid line), and (c) over estimated reverberant time (solid line). The dotted line indicates the power envelope of original clean speech.

reverberant environment," *Proc. ICSLP '06*, pp. 2546–2549 (2006).
[25] http://sp.shinshu-u.ac.jp/CENSREC/, AURORA-2J database.
[26] *The HTK Book* (*version 3.2*), Cambridge University Engineering Department (2002).
[27] X. Lu, M. Unoki and M. Akagi, "Sub-band temporal envelope restoration for ASR in reverberation environment," *IEICE Tech. Rep.*, SP2005-175, pp. 73–78 (2006).
[28] SMILE2004, *Sound Material in Living Environment*, Architectural Institute of Japan and GIHODO SHUPPAN Co., Ltd., (2004).
[29] B. J. Shannon and K. K. Paliwal, "A comparative study of filter bank spacing for speech recognition," *Proc. Microelectronic Engineering Research Conf.* (2003).

APPENDIX I

The effects of over- and under estimation of the reverberant time T_R on power envelope restoration is illustrated in Fig. 6, which has been taken from the 20th frequency channel of Fig. 2, as an example.

From Fig. 6, one can see that reverberant time $\hat{T}_{R,n}$ (0.31 s in this case) estimated using Eq. (10) can guarantee the best restoration of the power envelope, while under estimation (about half of $\hat{T}_{R,n}$: 0.16 s) and over estimation (about $1.5 \times \hat{T}_{R,n}$: 0.47 s) did not yield best restoration of the power envelope. In inverse filtering, our algorithm reveals the best $\hat{T}_{R,n}$ for restoring the power envelope so that the estimated values are not the same for all sub-bands. Moreover, most of the estimated values are not equal to the original reverberant values. In over estimates, the restored power envelope in each sub-band is high-pass filtered with a higher-end frequency than that used for accurate estimates, and vice versa for under-estimates. We explained the effect of power envelope restoration in Subsection 2.2.1. We tested its effect on ASR in over- and under-dereverberation, and presented the results in Fig. 7.

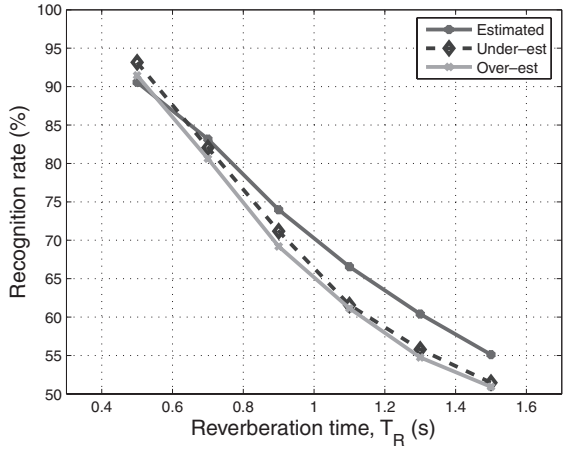

Fig. 7 Effects of over- and under estimated reverberant time T_R on recognition accuracy.

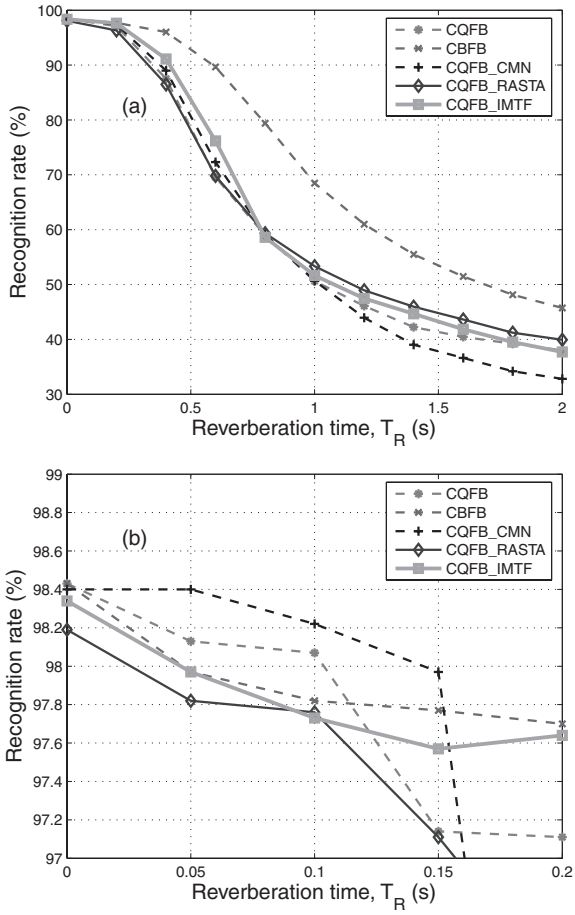

Fig. 8 Comparative evaluations of reverberant speech-recognition rates: (a) whole evaluation and (b) close-up of plot in range from 0.0 to 0.2 s.

From Fig. 7, we can see that the ASR system exhibits the best performance when the reverberant time is estimated using our new approach [27].

APPENDIX II

The accuracies of speech recognition when using CQFB, CBFB, CQFB_CMN, CQFB_RASTA, and CQFB_IMTF are plotted in Fig. 8. These are additional results to those in Fig. 5 and they have the same format. From the figure, we can see that CMN and RASTA processing on CQFB, on average, do not improve recognition any more than does the CQFB alone. Moreover, dereverberation based on the MTF concept in CQFB does not improve the performance because CQFB processing does not satisfy the requirements of MTF-based inverse filtering.