A robust feature extraction based on the MTF concept for speech recognition in reverberant environment

Xugang Lu, Masashi Unoki, and Masato Akagi

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

10.21437/Interspeech.2006-638

Abstract

This paper proposes a robust feature extraction method for automatic speech recognition (ASR) systems in reverberant environment. In this method, a sub-band power envelope inverse filtering algorithm based on the modulation transfer function (MTF), that we have previously proposed, is incorporated as a front-end processor for ASR. The impulse response of the room acoustics is assumed to be exponential decay modulated white noise, and speech is assumed to be temporal modulated white noise in each sub-band. Therefore, the impulse response of the environment does not need to be measured. Testing demonstrated that this algorithm can restore the temporal power envelope of reverberant speech in sub-bands and thus reduce the loss of speech intelligibility caused by reverberation. Testing of its ability to recognize digitized Japanese speech was done by using reverberant speech created by simple convolution of the room acoustics and speech. The algorithm had a 32.1% higher error reduction rate (on average, for reverberation times from 0.1 to 2.0 s) compared with the traditional cepstral mean normalization (CMN) of the auditory power spectrum based method (AFCC).

Index Terms: dereverberation, speech recognition, modulation transfer function

1. Introduction

Achieving robust speech recognition in reverberant environment is a big challenge in the speech recognition field. Reverberation can be regarded as convolution processing between acoustic speech and room acoustics. In a reverberant environment, the temporal and spectral structure of speech is distorted by stochastic reverberation caused by room reflection characteristics. It is difficult to distinguish clean speech signals in a reverberant environment by using the statistical properties of the speech and noise. Thus, the traditional noise reduction methods, such as spectral subtraction, Wiener filtering, and Bayesian estimation, which use different statistical properties of speech and noise, do not work well in reverberant environments. Several algorithms for reducing reverberation distortion have been proposed, e.g., cepstral mean normalization (CMN) [1] and RASTA filtering [2]. However, they are only suitable for short convolved noise or short reverberant situations. In actual room acoustics, reverberation time is far longer, and the properties of a reverberant environment are time-variant in a short time window. Several dereverberation algorithms using single- or multi-microphones have been proposed for solving the room reverberation problem. The basic principle of dereverberation is to estimate the impulse response of the room acoustics, and then use inverse filtering to obtain the dereverberated speech [3]. However, estimating the impulse response of room acoustics from only observed reverberant speech it is very difficult. One approach to speech dereverberation without estimating the impulse response of the room acoustics is to use speech characteristics. For example, the harmonic structure of speech can be used [4]. However, the harmonic structure is usually distorted in reverberant speech and hard to extract.

Speech signals are highly temporally modulated, and most of their intelligibility information is encoded in temporal modulation envelopes in each frequency band [5]. This means that, in the speech recognition task, we need only to restore the temporal envelope of clean speech from the reverberant speech in each frequency band. We previously proposed a sub-band power envelope inverse filtering algorithm based on the modulation transfer function (MTF) for dereverberating speech signals [6, 7]. It was designed to be used as a front-end processor for automatic speech recognition. Correlation and SNR measurements showed that it improves power envelope restoration accuracy [6, 7], and testing showed that it restores speech signals with a high level of speech intelligibility [8]. We have now tested its ability to recognize digitized Japanese speech.

2. MTF-based sub-band power envelope restoration

Before we discuss modeling the reverberant effect, we will briefly describe the MTF concept. The complex MTF is defined [9] as:

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

where $h(t)$ is the impulse response of the room acoustics, and ω is the radian frequency. For room acoustics, a well-known stochastic approximation of the impulse response is defined [9] as:

$$h(t) = e_h(t)\mathbf{n}_1(t) = a\exp(-6.9t/T_R)\mathbf{n}_1(t), \tag{2}$$

where $e_h(t)$ is the exponential decay temporal envelope, a is a constant amplitude, and $\mathbf{n}_1(t)$ is a random white noise. The corresponding MTF is obtained using:

$$m(\omega) = |\mathbf{M}(\omega)| = \left[1 + \left(\omega\frac{T_R}{13.8}\right)^2\right]^{-1/2}. \tag{3}$$

In Eqs. (2) and (3), T_R is the reverberant time defined as the time required for the power of $h(t)$ to decay by 60 dB [6]. For a dominant frequency in the temporal envelope, Eq. (3) can be regarded

as the modulation index, i.e., the degree of the relative fluctuation in the normalized amplitude with respect to the dominant frequency. On the basis of this characteristic, T_R can be predicted from a specific frequency by using the MTF.

We model the effect of room acoustics on speech signals based on the MTF concept. The convolution distortion in sub-band representation is written as

$$y_n(t) = x_n(t) * h(t), \qquad n = 1, 2, \cdots, N, \tag{4}$$

where $y_n(t)$ and $x_n(t)$ are 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 property of the speech signal, we model the sub-band speech, $x_n(t)$, as

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

The temporal envelope of sub-band n is $e_{x,n}(t)$. In Eqs. (2) and (5), $\mathbf{n}_1(t)$ is mutually independent white noise that satisfies

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

where $< \cdot >$ is the ensemble average operator. Using Eqs. (4) - (6), the power envelope of $y_n(t)$ can be calculated [6] as:

$$\left\langle y_n(t)^2 \right\rangle = e_{y,n}(t)^2 = e_{x,n}(t)^2 * e_h(t)^2. \tag{7}$$

This equation shows that the restoration of $e_{x,n}(t)^2$ can be completed by deconvolution of $e_{y,n}(t)^2$ with $e_h(t)^2$. For discrete signals in z-transforms, the deconvolution is

$$E_{x,n}(z) = \frac{E_{y,n}(z)}{E_h(z)}, \tag{8}$$

where $E_{x,n}(z)$, $E_{y,n}(z)$, and $E_h(z)$ are the z-transforms of $e_{x,n}(t)^2$, $e_{y,n}(t)^2$, and $e_h(t)^2$, respectively. Thus,

$$E_h(z) = \frac{a^2}{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}}. \tag{9}$$

Substituting Eq. (9) into Eq. (8), we get

$$E_{x,n}(z) = \frac{E_{y,n}(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\}. \tag{10}$$

The power envelope of sub-band signal $e_{x,n}(t)^2$ can be restored using the inverse z-transform of $E_{x,n}(z)$. In Eq. (10), we need to estimate parameters a and T_R.

3. Algorithm implementation

Our sub-band power envelope inverse filtering algorithm was developed on the basis of the analysis above. As shown in Fig. 1, the observed signal $y(t)$ is decomposed into a series of frequency sub-bands; envelope detectors then extract temporal modulation envelopes $e_{y,n}(t)^2$. Given the co-modulation characteristics of speech signals in sub-bands [7], we use a series of FIR-type band-pass filters with a constant bandwidth of 100 Hz for the decomposition. The extracted envelopes are used for inverse filtering, which is controlled by estimated parameters $\hat{a}$ and $\hat{T}_R$. The final output is the restored or dereverberated power envelope, $\hat{e}_{x,n}(t)^2$, for each frequency band. The power envelope inverse filtering is done for each sub-band in three steps.

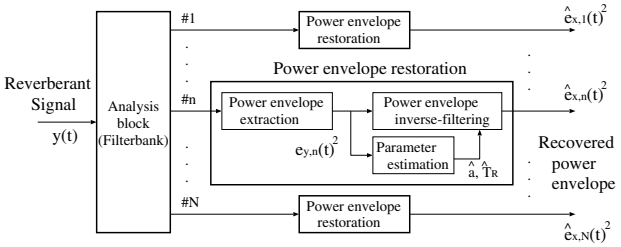

Figure 1: Sub-band power envelope inverse filtering algorithm.

3.1. Extract sub-band power envelopes

Low-pass filtering with half-wave rectification (HWR) is widely used for sub-band temporal envelope estimation in computational auditory models. However, since we assume the carrier in each sub-band is white noise rather than a monotone sine wave, we extract the power envelopes using low-pass filtering of the Hilbert transform of the sub-band signals [6, 7]:

$$\hat{e}_{y,n}(t)^2 = \text{LPF}\left[\left|y_n(t) + j\text{Hilbert}(y_n(t))\right|^2\right], \tag{11}$$

where **LPF**$[\cdot]$ is a low-pass filtering operator, and **Hilbert**$(\cdot)$ is the Hilbert transform. We set the cut-off frequency of the low pass filtering to 20 Hz in order to keep most of the important modulation information for speech perception.

3.2. estimate parameters of room acoustics

We estimate parameters T_R and a. In Eq. (**??**), [6] using

$$\hat{T}_R = \max\left(\underset{T_{R,\min}<T_R<T_{R,\max}}{\arg\min} \int_0^T \left|\min\left(\hat{e}_{x,n,T_R}(t)^2, 0\right)\right| dt\right), \tag{12}$$

and

$$\hat{a} = \sqrt{1/\int_0^T \exp\left(-\frac{13.8t}{\hat{T}_R}\right) dt} \tag{13}$$

where T is signal duration and $\hat{e}_{x,n,T_R}(t)^2$ represents the candidates of the restored power envelope as a function of T_R. The $T_{R,\min}$ and $T_{R,\max}$ are the lower and upper bounds of T_R. Equations (12) and (13) are described in detail elsewhere [6, 7].

3.3. Inverse filter of power envelopes

After the power envelopes ($e_{y,n}(t)^2$) and the parameters of the room acoustics ($\hat{T}_R$ and $\hat{a}$) are obtained, the power envelopes are inverse filtered using Eq. (10) to restore the power envelopes of the dereverberated speech in the sub-bands. The effects of the algorithm are illustrated in Fig 2, which shows the processing of digitized Japanese speech in three sub-bands with center frequencies of 1.0, 2.0, and 3.0 kHz.

The solid curves show the sub-band power envelopes of clean speech in the three sub-bands, while the dashed curves in the left panels show the power envelopes of reverberant speech (with $T_R = 0.7$ s) extracted without any dereverberation processing. The sub-band power envelopes of the reverberant signal diffuse from the enveloped peaks with an exponential decay that distorts the subsequent temporal envelopes of the signal. The dashed curves in the right panels show the power envelopes extracted using the proposed dereverberation processing. In each sub-band,

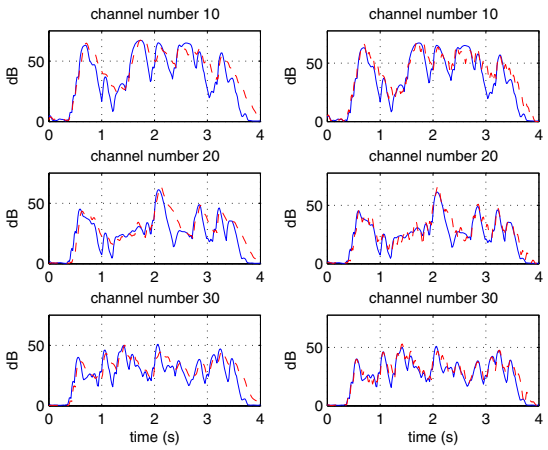

Figure 2: Sub-band power envelope of clean and reverberant speech without (left) and with (right) dereverberation processing.

the restored power envelope is closer to the power envelope of the clean speech signal. Consequently, if the restored sub-band power envelope is used to extract speech features for speech recognition, the recognition of reverberant speech should be improved.

4. Speech recognition for reverberant speech

We used the dereverberation algorithm as a front-end processor for automatic speech recognition (ASR) to test its recognition of reverberant speech. We used clean speech from AURORA-2J as speech material [10] and used 8840 clean speech sentences to train the acoustic models. For testing, we used 1001 reverberant speech sentences produced artificially by convolving the speech signals with a room acoustic impulse response signal with a reverberation time of 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9 or 2.0 s. The sampling frequency, f_s, was 8 kHz, so we used 40 sub-band channels ($N = 40$) to cover the frequency region from 0 to 4 kHz.

The speech features were extracted using the restoration process described above, as Illustrated in Fig. 3. The input used to restore the power envelopes of the clean speech in each sub-band were smoothing blocks comprising frame integration and log compression. Because the power envelope inverse filtering is high-pass, low-pass filtering with forgotten parameter λ was used for smoothing the envelope dips in each sub-band:

$$\hat{I}_i(t) = \lambda \hat{I}_i(t-1) + (1-\lambda) I_i(t), \qquad (14)$$

where $I_i(t)$ is the original restored sub-band power envelope, and $\hat{I}_i(t)$ is the smoothed output. We set λ to 0.99. For the frame integration, 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 done. The DCT was used for dimensional decorrelation. The first 12 dimensions of the decorrelated log power spectrum were used (the zero-th order coefficient was discarded). Combining the log power energy, we obtained 13-dimension static feature sets. Together with their first and second order delta dynamic values, 39 dimensions feature vectors were formed. HTK [1] was used for training the HMM acoustic models. The acoustic models were configured the same as in

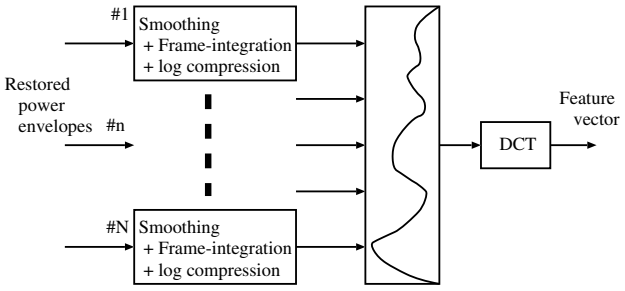

Figure 3: Speech feature extraction based on restored power envelope in each sub-band.

the AURORA-2J experiments. We investigated the effects of some processing methods on ASR performance [10].

4.1. Effect of band-pass filtering bandwidth

Mel band-pass filtering based feature extraction is widely accepted as being more robust in most additive nose conditions than constant band-pass filtering. However, because our MTF-based dereverberation is based on sub-band power envelope inverse processing, the power envelope should have a co-modulation characteristic in one sub-band while satisfying the MTF concept. We carefully choose the bandwidth by considering the trade-off between them. For an initial experiment, we tested the effect of using equivalent rectangular bandwidths (ERBs) of gammatone filters and constant filter bandwidths on recognition performance. We found that a constant bandwidth of 100 Hz is more suitable for satisfying the envelope co-modulation property and the MTF concept. Therefore, in our MTF-based dereverberation experiments, we used band-pass filters with a 100 Hz bandwidth.

4.2. Effect of over- and under-dereverberation

In the inverse filtering, the estimation of parameter T_R for the room acoustics is important. Our algorithm determines the T_R that is best for restoring the power envelope and so that the estimated values are not the same in all sub-bands. Moreover, most of the estimated values are not equal to the original value. In the case of over-estimation, the restored power envelope in each sub-band is high-pass filtered with a higher end frequency than used for an accurate estimation, and vice versa in the case of under-estimation. We tested the effect on ASR of over and under-dereverberation and found that both over- and under-dereverberation reduce recognition accuracy compared with our estimated $\hat{T}_R$ because they are not appropriate for achieving the best restoration.

4.3. Comparison with traditional feature extraction methods

We used RASTA filtering [2] of the auditory power spectrum based speech feature extraction and CMN of the auditory power spectrum based cepstral coefficients for comparison purposes. The results are shown in Fig. 4, where "AFCC-RASTA" represents cepstral feature extraction is based on RASTA filtering on the log auditory power spectrum of each sub-band, and "AFCC-CMN" represents CMN of the auditory power spectral based cepstral coefficient. The auditory power spectrum was calculated using gammatone band filters with ERBs and half-wave rectifying in each sub-band. "No processing" represents using the proposed constant-

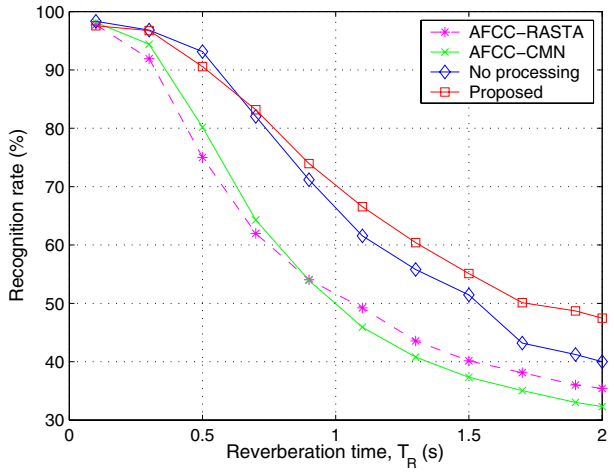

Figure 4: Comparison of reverberant speech recognition rates.

bandwidth band-pass filtering and power envelope extraction without dereverberation. "Proposed" represents cepstral feature extraction based on our proposed MTF-based sub-band power envelope estimation with dereverberation.

As shown in Fig. 4, the speech recognition rate decreased as the reverberation time was increased; the rate of decrease was especially high when the reverberation time was long ($T_R > 0.5$ s). When it was ($T_R < 0.3$ s), all the feature extraction methods performed well (recognition rate $> 90\%$). However, for reverberation times of 0.5 to 2.0 s, the recognition rates of AFCC-RASTA and AFCC-CMN decreased more sharply than those of our proposed methods. On average, for reverberation times from 0.1 s to 2.0 s, the proposed MTF-based dereverberation algorithm had a 32.1 % better error reduction rate that AFCC-CMN; it was 9.86 % better the proposed band-pass filtering and power envelope estimation without dereverberation processing (No processing).

5. Discussion and conclusion

Our analysis and experiments demonstrated that our MTF-based sub-band power envelope inverse filtering algorithm improves the robustness of speech recognition performance for reverberant speech, especially for long reverberant situations. The results showed: (1) MTF-based dereverberation can restore the sub-band temporal power envelope of speech, thereby improving automatic speech recognition performance for reverberant speech; (2) band-pass filtering with a constant bandwidth of 100 Hz is better than that with ERB for dereverberation; and (3) under- and over-dereverberation based feature extraction both degrade ASR performance.

Comparison of the recognition rates when the proposed method was used with that when no processing was used showed that the recognition rate was still relatively low. This suggests that we need to reconsider how some things are handled. For example, dereverberation is done using the estimated reverberation time in each sub-band independently. If there is even a small error in the estimations, the extracted feature may differ greatly from the actual feature due to temporal misalignment between sub-bands. A more accurate way is thus need for estimating reverberation time. We also need to find a more accurate method of estimating the sub-band temporal power envelopes because the inverse filtering for dereverberation is based on these envelopes. We need a way to estimate the sub-band temporal power envelope based on stochastic signal processing for both Gaussian and non-Gaussian white-noise carriers. Finally, our experiments were based on artificial reverberant speech. We plan to record a speech data corpus in an actual reverberant environment and use if for testing.

6. Acknowledgements

This work was supported by a Grant-in-Aid for Science Research from the Ministry of Education (No. 18680017). We would like to thank ATR Spoken Language Translation Research Laboratories for permitting us to use the AURORA-2J data.

7. References

[1] The HTK Book (version 3.2), Cambridge University Engineering Department, 2002.

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

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. Acoustic. Speech signal process., **ASSP-36**, 145–152, 1988.

[4] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," Proc. ICASSP 2003, **1**, 92–95, 2003.

[5] R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech Recognition with Primarily Temporal Cues," Science, **270**, 303–304, 1995.

[6] 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.

[7] 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**(4), 243–254, 2004.

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

[9] M. R. Schroeder, "Modulation transfer function: definition and measurement," Acustica, **49**, 179–182, 1981.

[10] http://sp.shinshu-u.ac.jp/CENSREC/,AURORA-2J database.