INTERSPEECH 2009 BRIGHTON

Subband Temporal Modulation Spectrum Normalization for Automatic Speech Recognition in Reverberant Environments

Xugang Lu [1], *Masashi Unoki* [2], *Satoshi Nakamura* [1]

[1]National Institute of Information and Communications Technology, Japan
[2]Japan Advanced Institute of Science and Technology, Japan

Abstract

Speech recognition in reverberant environments is still a challenge problem. In this paper, we first investigated the reverberation effect on subband temporal envelopes by using the modulation transfer function (MTF). Based on the investigation, we proposed an algorithm which normalizes the subband temporal modulation spectrum (TMS) to reduce the diffusion effect of the reverberation. During the normalization, both the subband TMS of the clean and reverberated speech are normalized to a reference TMS calculated from a clean speech data set for each frequency subband. Based on the normalized subband TMS, the inverse Fourier transform was done to restore the subband temporal envelopes by keeping their original phase information. We tested our algorithm on reverberated speech recognition tasks (in a reverberant room). For comparison, the traditional Mel-frequency cepstral coefficient (MFCC) and relative spectral filtering (RASTA) were used. Experimental results showed that the recognition rate using the feature extracted based on the proposed normalization method has totally a 80.64% relative improvement.

Index Terms: Dereverberation, temporal modulation, sub-band temporal envelope, automatic speech recognition.

10.21437/Interspeech.2009-374

1. Introduction

In reverberant environments, the reverberation decreases the intelligibility of speech perception as well as the performance of automatic speech recognition (ASR) systems. Although, there are many classical and successful noise reduction methods, however, most of them are based on modeling the difference of the statistical properties of noise and speech. The statistical characteristic of the reverberated speech is the same as that of clean speech, therefore it is difficult to reduce the reverberation effect using those traditional methods.

The basic principle of dereverberation is to measure the impulse responses (IRs) of room acoustics or propagation channels, and then use inverse filtering to obtain the dereverberated speech [1]. However, those methods require the IRs of room acoustics for each dereverberation process to be remeasured if the conditions for room acoustics change. Blind dereverberation, which does not need the estimation of the IRs of room acoustics, is preferred for real applications. One possible way in blind dereverberation is to use the speech characteristics. For example, the harmonic structure of speech can be used [2]. This method needs the fundamental frequency from reverberated speech to be accurately estimated, which is difficult, and it does not seem to restore the consonant (nonharmonic) parts in speech.

In this study, we mainly focus on the robustness problem for ASR in reverberant environments. Therefore, we only need to reduce the reverberation effect on the speech representations rather than to recover the original speech waveform. As many studies showed that speech signals are highly temporally modulated in amplitude, and most of their intelligibility information is encoded in the temporal modulation envelopes of several frequency subbands [3, 4]. In a reverberant room, the reverberation effect on the temporal modulation envelopes has been investigated for many years. It was found that in order to keep enough intelligibility, the temporal modulation envelopes of speech should be kept as much as possible from one spatial location to another location during the communications in a room. Based on the temporal modulation property of speech, several algorithms were proposed to normalize the temporal modulation spectrum of speech for noisy speech processing. However, most of them dealt with additive noise and short-term convolution effect, and did the processing in cepstral domain [4, 9]. In their algorithms, the physical concept of the temporal modulation processing was not used accurately since the processing was done in cepstral domain. In addition, the reverberation effect in cepstral domain is spread to all dimensions of the feature, while in real situation, the reverberation effect is frequency-dependent. It is better to reduce the reverberation effect in temporal envelope domain before it is transformed to cepstral domain.

The reverberation effect on the temporal modulation envelopes has been investigated under the concept of modulation transfer function (MTF) [5]. In the concept underlying MTF, the transfer function of the room was found to be functioned as a low-pass filtering on the temporal modulation envelopes which is closely related with the reverberant time (RT) and speaker to microphone distance (SMD) [6]. Based on the MTF concept, an inverse MTF (IMTF) method was developed for dereverberation based on temporal envelope inverse filtering [7]. This method was also used for speech feature extraction and tested on reverberated speech recognition [8]. However the IMTF is a model based dereverberation which gives several assumptions during the model derivation. Since the clean and reverberated speech do not fit to the model assumptions well, the improvement of the performance was limited. Rather than using signal models for dereverberation in temporal envelope modulation domain, in this paper, we propose to normalize the modulation spectrum of the subband temporal envelopes for the speech signal to reduce the reverberation effect.

2. Reverberation effect on temporal envelopes

We analyze the reverberation effect by using the temporal modulation contrast changes which can be quantified by using the temporal modulation spectrum (TMS) (Fourier transforms of

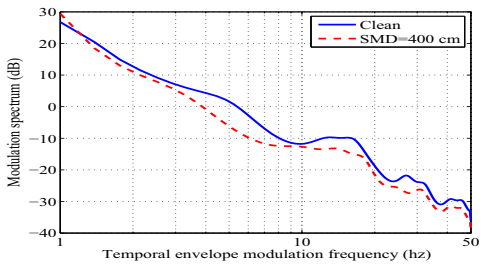

Figure 1: Temporal modulation spectrum of clean (solid) and reverberated (dashed) speech (with center frequency 1kHz).

the temporal modulation envelopes that measure the ensemble of the temporal modulation changes due to reverberation). Suppose $a_x(t)$ and $a_y(t)$ are the subband temporal envelopes (STEs) of the clean and reverberated speech, and their Fourier transforms are $A_x(\Omega)$ and $A_y(\Omega)$, respectively (for convenience of discussion, we omit the subband index in this paper). The Ω is the modulation frequency. The average or ensemble of the STEs can be quantified using the average of the power spectral density (PSD) of the STE as:

$$P_{xx}(\Omega) = \left\langle \frac{|A_x(\Omega)|^2}{N} \right\rangle ; P_{yy}(\Omega) = \left\langle \frac{|A_y(\Omega)|^2}{N} \right\rangle, \quad (1)$$

where N is the length of the STE and $\langle . \rangle$ is the ensemble average operator.

Because the reverberation effect depends on the linguistic context of speech utterances, we consider the reverberation effect by using the smoothed temporal modulation PSD to smooth out the details of the PSD due to linguistic context effect. An example of the smoothed PSD of the STE of a clean and reverberant speech (with SMD=400 cm) is shown in Fig. 1. From this figure, we can confirm that the modulation spectrum of the reverberated speech can be regarded as a low-pass filtered one of the clean speech. This low-pass filter is the subband temporal modulation transfer function (STMTF) between the clean and reverberant speech. This transfer effect of the room acoustic causes the decrease of the speech intelligibility [6]. In order to reduce this effect, we need to do the inverse filtering to normalize the modulation spectrum of the reverberated speech to that of the clean speech. Based on the normalization, the clean temporal modulation structure of speech can be restored.

3. Proposed temporal modulation spectrum normalization algorithm

If we know all the STMTFs, we can do the inverse filtering to recover the original STEs. However, in real applications, the clean speech utterances corresponding to the reverberated ones are unknown, we only have the observed reverberated utterances. It is difficult to estimate the STMTFs or the inverse filters. Rather than recovering the original STE of the clean speech directly, we attempt to normalize the subband temporal modulation PSDs of the clean and reverberated speech to a reference modulation PSD. Therefore the transform relationship between the clean, reverberated and normalized PSDs of the temporal modulation can be shown as in Fig. 2. In this figure, both the modulation PSDs of the clean and reverberant speech can be normalized to a reference modulation PSD $P_{rr}(\Omega)$ via the transforms $H_{xr}(\Omega)$ and $H_{yr}(\Omega)$, respectively. The PSD

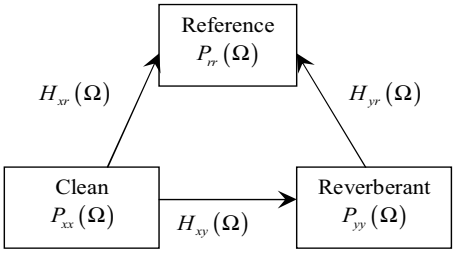

Figure 2: Normalization of the modulation power spectrum.

of the STMTFs between the clean and reverberated speech is $H_{xy}(\Omega) = H_{xr}(\Omega)H_{yr}^{-1}(\Omega)$. This kind of normalization can also be regarded as the modulation spectrum representation in different coordinative systems that are determined by the environments. The $H_{xr}(\Omega)$ and $H_{yr}(\Omega)$ can be regarded as the PSDs of the modulation transfer filters between the clean, reverberated and the reference environments, respectively. In our study, the reference PSD is set as the average PSD of the STE of a prior clean speech data set in each frequency band as:

$$P_{rr}(\Omega) = \frac{1}{M}\sum_{i=1}^{M} P_{xx}^{i}(\Omega), \quad (2)$$

where $P_{xx}^{i}(\Omega)$ is the modulation PSD of the i-th clean utterance, $i = 1, 2, \cdots, M$, with M is the total number of the speech utterances. Because the transfer function is different utterance by utterance, we estimate the transfer functions using the following equations as:

$$H_{xr}(\Omega) = \frac{P_{rr}(\Omega)}{P_{xx}(\Omega)}; H_{yr}(\Omega) = \frac{P_{rr}(\Omega)}{P_{yy}(\Omega)}. \quad (3)$$

Based on these transforms, we can see that the transfer function keeps not only the transform relationship between different environments, but also the utterance dependency property which retains the discriminative information due to the difference of the temporal modulation.

Based on the normalized modulation PSD, we can obtain the normalized temporal envelopes of the clean and reverberated speech as:

$$\hat{a}_x(t) = \text{Real}\left(\text{IFFT}\left(\hat{A}_x(\Omega)\right)\right), \quad (4a)$$

$$\hat{a}_y(t) = \text{Real}\left(\text{IFFT}\left(\hat{A}_y(\Omega)\right)\right), \quad (4b)$$

where

$$\hat{A}_x(\Omega) = |A_x(\Omega)|\sqrt{H_{xr}(\Omega)}\exp\left(j\theta_x(\Omega)\right), \quad (5a)$$

$$\hat{A}_y(\Omega) = |A_y(\Omega)|\sqrt{H_{yr}(\Omega)}\exp\left(j\theta_y(\Omega)\right), \quad (5b)$$

$$\theta_x(\Omega) = \arg\left(A_x(\Omega)\right); \theta_y(\Omega) = \arg\left(A_y(\Omega)\right). \quad (5c)$$

Here, 'Real' and 'IFFT' represent the real part of the inverse fast Fourier transform, respectively. In the estimation of the PSD of the STE of the clean speech in Eqs. 1 and 2, the smoothed PSD is used. By this processing, the reverberation effect is removed while the discriminative information of each utterance due to the difference of modulation is kept.

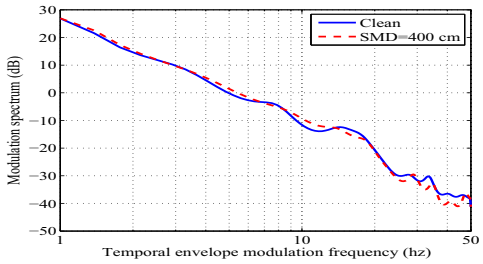

Figure 3: Normalized subband temporal modulation PSDs of the clean (solid) and reverberated (dashed) speech.

4. Experimental evaluations

In this section, we evaluate the proposed method for temporal envelope restoration of reverberated speech, and test the robustness of the feature extracted based on the restored envelopes on reverberated speech recognition experiments. Before doing the modulation spectrum normalization, the observed speech signal is first decomposed into several frequency subbands. Considering the subband temporal envelope co-modulation property of speech, a series of FIR-type band pass filters with constant bandwidth (40 band-pass filters with 100 Hz bandwidth to cover the frequency range from 0 Hz to 4000 Hz in this study) are designed. The temporal envelope in each frequency subband is estimated using a low-pass filtering of the Hilbert transform of the subband signal [7].

4.1. Normalized subband temporal modulation spectrum and restored temporal envelopes

After getting the STEs, we first calculate the average temporal modulation PSD of clean speech in each frequency subband using Eq. 2 (one clean training data set from AURORA-2J data corpus is used in this study [10]). The average temporal modulation PSD is then used as a reference to normalize both the temporal modulation PSDs of the clean and reverberated speech in each frequency band. Based on the normalized temporal modulation PSD, the STEs are recovered using Eqs. 4 and 5.

An example is shown in Fig. 3 for the normalized subband temporal modulation PSD (the same utterance and the same frequency band as used in Fig. 1). Comparing Figs. 3 with 1, we can see that the normalized modulation PSDs of the clean and reverberated speech match each other better. Based on these normalized PSDs, the STEs are restored. An example of the STEs of an utterance of the original reverberated speech and normalized one are shown in Fig. 4. From this figure, we can see that the normalization processing is similar as a high-pass filtering to enhance the speech temporal modulation events which is a function of the inverse filtering of the STMTFs. Based on this normalization, the reverberation effect is reduced. In addition, because the average clean modulation PSDs are used as references, the mismatch between the normalized clean and normalized reverberated speech is reduced (the normalized spectrum of clean speech is not given here because of the limited space). After getting the restored STEs, we extract the the cepstral feature, and test the robustness for ASR.

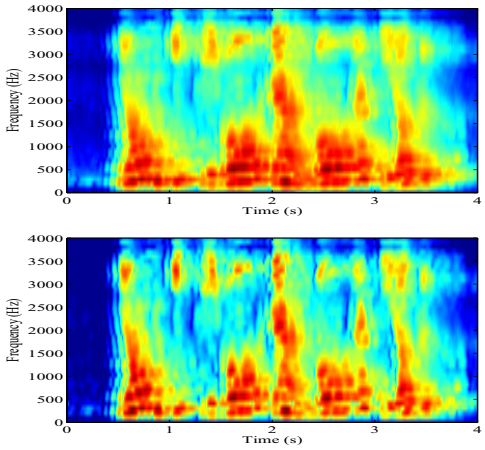

Figure 4: Subband STEs based spectrum of the original reverberated speech (upper panel) and the restored one (lower panel).

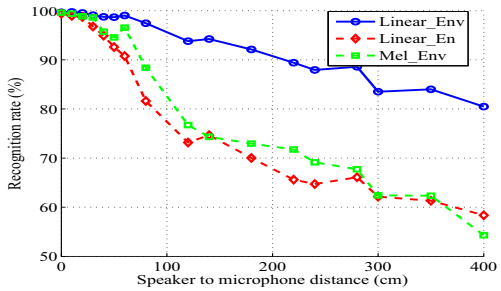

Figure 5: Recognition performance for constant and nonlinear bandwidth subband processing with subband energy and temporal envelope as outputs.

4.2. Speech recognition experimental conditions

In our speech recognition experiments, clean speech from the AURORA-2J database was used [10]. 8840 clean speech utterances were used to train the acoustic models. 1001 clean speech utterances convolved with the impulse response of a reverberant room was used for testing. The acoustic models, and the configuration of feature calculation were set the same as those used in the AURORA2J experiments [10]. The HTK speech toolkit was used for training the HMM acoustic models. Before comparing our proposed method with traditional methods, we compare the recognition performances of using the subband temporal envelope and power energy based features. All the features are the decorrelated log compressed subband output (40 filter bands are used) via a discrete cosine transform (DCT) to get the cepstral coefficients.

4.3. Subband temporal envelope vs power energy

In this study, we have tested that the selection of subband filters does not have much effect for reverberated speech if the subband energy is used as the output. However, considering the subband temporal envelope comodulation property of speech, when using the temporal envelope as the subband output, our specially designed constant bandwidth FIR filters are more suitable. The performance of using the two types of subband outputs in feature extraction for ASR is showed in Fig. 5. In this

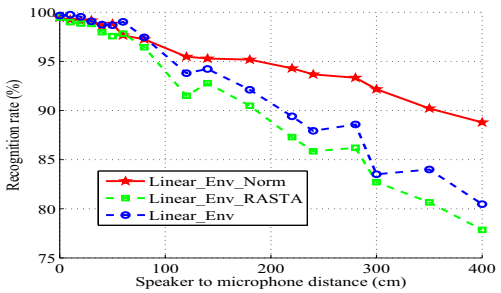

Figure 6: Recognition performance for filtering on the subband temporal envelope.

figure, the 'Linear_Env' and 'Linear_En' represent the temporal envelope and energy as the subband outputs for the constant bandwidth filters, respectively. The 'Mel_Env' represents the temporal envelope as the subband output for the Mel bandwidth filters. From Fig. 5, we can see that if the subband temporal envelope is used for feature extraction, for Mel bandwidth subband processing, there is only a little improvement. However, using our designed constant bandwidth subband processing, the performance is improved significantly. Based on these results, we can find that the subband temporal envelope is more robust for reverberated speech processing since the temporal envelope is usually dominated by the modulation events with large energy peaks which are not easily masked in reverberant environments. In the Mel bandwidth subband processing, because different temporal modulation envelopes are possibly mixed in one subband, the modulation events with small temporal envelope peaks may be masked by the reverberation effect of the modulation events with large temporal envelope peaks. Therefore, the recognition performance is decreased.

4.4. Normalization on subband temporal modulation spectrum vs RASTA filtering

Our proposed method is functionally equal to a filtering on the subband temporal envelope using a specially designed filter which equalizes the subband temporal modulation spectrum. From this point of view, it is similar as the RASTA processing on subband temporal envelope. Therefore, we choose the RASTA filtering as the traditional method for comparison. The comparison is shown in Fig. 6. In this figure, 'Linear_Env_Norm' and 'Linear_Env_RASTA' represent the normalization on temporal modulation spectrum and RASTA filtering methods on the temporal envelope of the constant bandwidth filters, respectively. From Fig. 6, we can see that when the SMD is short (less than 80 cm in this case), there is almost no effect for the proposed and RASTA filtering methods or a little decrease in recognition performance. However, when the SMD is long (larger than 80 cm in this case), the proposed method improved the performance significantly, while the RASTA filtering method surprisingly decreased the performance. This result confirms that the filtering of the temporal envelope based on the normalization of the temporal modulation spectrum can reduce the reverberation effect, while the RASTA filtering does not.

5. Conclusion and discussions

In this paper, we fist investigated the reverberation effect on the subband temporal envelopes. Based on the investigation, we found that the modulation contrast and modulation spectrum of the temporal envelopes are systematically changed in reverberant environments which is well explained using the MTF concept of the room acoustic. Under this concept, we proposed to normalize the subband temporal modulation spectrum to reduce the reverberation effect. Based on the normalized modulation spectrum, the subband temporal envelopes are restored which is used in speech feature extraction for ASR. Experimental results showed that there is a significant improvement by using the extracted feature compared with the traditional features, i.e., by averaging the SMDs from 80 cm to 400 cm, there is 65.87% relative improvement by using only the constant subband temporal envelope processing compared with MFCC, and a further 34.59% relative improvement by using the normalization on the subband temporal modulation spectrum. Totally, there is about a 80.64% relative improvement in recognition rate.

In the proposed method, the average TMS of the clean speech data set was used as a reference. However, the averaged TMS is different from those of the TMS for each speech utterance. It is better to use an adaptive TMS normalization method to recover the temporal modulations of speech utterance by utterance. Finding an adaptive TMS normalization for subband temporal envelopes is our future work.

6. Acknowledgements

This study is supported by the MASTAR project of Knowledge Creating Communication Research Center of NICT.

7. References

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. on Acoustics, speech, and signal processing, ASSP (36), 145-152, 1988.

[2] T. Nakatani, M. Miyoshi and K. Kinoshita, "Blind dereverberation of monaural speech signals based on harmonic structure," IEICE D-II, J88-D-II(3), 509-520, 2005.

[3] Drullman, R., Festen, J. M., Plomp, R., "Effects of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am., 95(5), 2670-2680, 1994.

[4] Kanedera, N., Arai, T., Hermansky, H., Pavel, M., "On the relative importance of various components of the modulation spectrum for automatic speech recognition," Speech Communication, 28 (1), pp. 43-55, 1999.

[5] 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.

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

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

[9] X. Xiao, E. S. Chng, and H. Li, "Normalization of speech modulation spectra for robust speech recognition," IEEE Trans. on Audio, Speech, and Language Processing, 16 (8), 1662-1674, 2008.

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