TEMPORAL CONTRAST NORMALIZATION AND EDGE-PRESERVED SMOOTHING ON TEMPORAL MODULATION STRUCTURE FOR ROBUST SPEECH RECOGNITION

X. Lu[1,2], *S. Matsuda*[1,2], *M. Unoki*[3], *T. Shimizu*[1,2], *and S. Nakamura*[1,2]

1. ATR-SLC. 2. National Institute of Information and Communications Technology, Japan.
3. Japan Advanced Institute of Science and Technology, Japan

ABSTRACT

In this paper, we propose a two-step processing algorithm which adaptively normalizes the temporal modulation of speech to extract robust speech feature for automatic speech recognition systems. The first step processing is to normalize the temporal modulation contrast (TMC) of the cepstral time series for both clean and noisy speech. The second step processing is to smooth the normalized temporal modulation structure to reduce the artifacts due to noise while preserving the speech modulation events (edges). We tested our algorithm on speech recognition experiments in additive noise condition (AURORA-2J data corpus), reverberant noise condition (convolution of clean speech utterances from AURORA-2J with a smart room impulse response), and noisy condition with both reverberant and additive noise (air conditioner noise in a smart room). For comparison, the ETSI advanced front-end (AFE) algorithm was used. Our results showed that the algorithm provided: (1) for additive noise condition, 57.26% relative word error reduction (RWER) rate for clean conditional training (59.37% for AFE), and 33.52% RWER rate for multi-conditional training (35.77% for AFE), (2) for reverberant condition, 51.28% RWER rate (10.17% for AFE) and (3) for noisy condition with both reverberant and additive noise, 71.74% RWER rate (48.86% for AFE).

Index Terms— robust speech recognition, temporal modulation contrast normalization, cepstral mean and variance normalization, modulation transfer function.

1. INTRODUCTION

In order to improve the noise robustness of automatic speech recognition (ASR) systems, many methods had been proposed, for example, spectral subtraction, Wiener filtering, etc. [1]. Most of them focus on reducing the noise effect in spectral domain. This strategy is based on the findings of noise effect on speech spectrum in a short-term period of speech samples (10 ms to 40 ms time window). However, as many researches showed that in the spectral domain speech spectrum is easily distorted either by additive noise or convolutive (reverberant) noise. In pursuing to find robust representations, many studies are done to find which information is essential for speech perception and robust to noise distortion. Recently, more and more evidences show that the temporal modulation structures (TMSs) are important for speech perception and relatively robust to noise environments [2, 3, 4]. Many experiments of speech perception show that the temporal modulation is an important cue for speech intelligibility. Drullman et al. carried out perception experiments and showed that most of the intelligibility information of speech is under 16 Hz [2]. Moreover, as showed by Shannon et al., if the TMSs are kept well with spectral information greatly reduced, the speech still keeps high intelligibility [3]. Based on these knowledge, some speech feature extraction methods were proposed. For example, relative spectra (RASTA) filtering [5], cepstral mean normalization (CMN), temporal trajectory filtering. The aim of those methods is to enhance the TMSs of speech. However, most of them usually adopt an average modulation information of speech for designing the modulation filters. For modulation information of speech which corresponds to the dynamic movement of articulators and multi-scale temporal organization of utterances actually distributes in a large range of modulation frequencies. For example, the TMSs of consonants and vowels are different, and the time scale of TMSs of syllables and phonemes are different. Therefore, it is better to find out an adaptive modulation filtering strategy to enhance the modulation information of speech with different local and global TMSs.

The purpose of this paper is to investigate the noise effects on modulation spectrum of speech in which the noise effect is regarded as modulation transfer function (MTF) either for additive noise or reverberant noise. Based on the investigation, we propose an adaptive speech processing strategy to extract robust speech features for ASR systems.

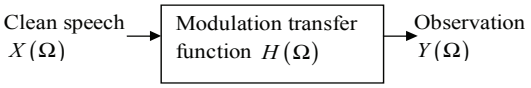

Fig. 1. Noise effect as modulation transfer function.

2. NOISE EFFECT ON TEMPORAL MODULATION OF SPEECH

In noisy environments, either reverberant noise or additive noise, the clean speech signal (speaker) is transmitted to the microphone (receiver) via a transmission environment, in terms of the transmission system concept, the transfer relation between clean speech and observed noisy observation is shown in Fig. 1.

In Fig. 1, Ω is the modulation frequency. $X(\Omega)$ and $Y(\Omega)$ are the modulation spectra of clean and observed speech signals. In this sense, the noise effect can be regarded as an MTF $H(\Omega)$. The MTF can be used as a low pass filter, high pass filter or a noise-dependent complex filter. The original MTF concept is referred to the temporal envelope modulation (usually in each frequency sub-band), but it can be the transformed TMS based on the temporal envelope modulation. Given the fact that most current ASRs use the static and dynamic properties of the mel frequency cepstral coefficients (MFCCs) as speech feature, we investigate the noise effect on temporal modulation of the cepstral coefficient, and the temporal modulation is referred as the cepstral TMSs hereafter.

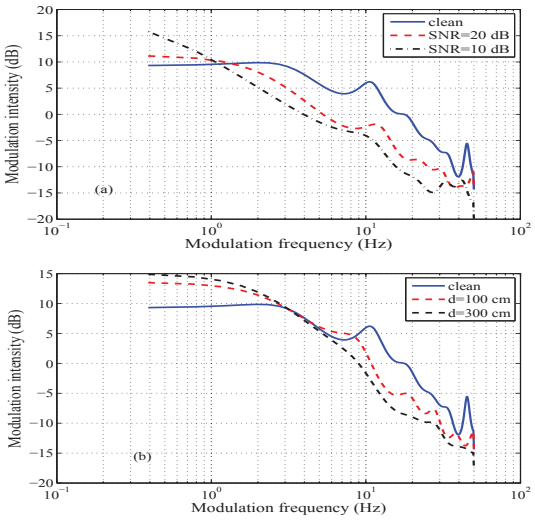

Fig. 2. Modulation spectrum in (a) additive noise and (b) reverberant noise conditions.

In additive noise condition, the noise energy may mask the speech energy, especially in the time locations of low speech energy. Correspondingly, in the cepstral domain, the temporal modulation contrast (TMC) may be changed, and the low or high temporal modulation fluctuations may be attenuated or enhanced. In reverberant condition, the effect of reverberant noise is to diffuse the high energy peaks of speech to the later speech components. If the diffusion energy is higher than the later speech energy, the later speech components will be masked. The effect of the reverberation depends on the reverberant time (RT) of the room and speaker to microphone distance (SMD). Generally speaking, the transfer function of the additive noise transfer system has low-pass filtering property which is closely related with the signal to noise ratio (SNR), while the transfer function of the reverberant room can be regarded as a low pass filtering on the temporal envelope structure which is closely related with the RT and SMD [6].

For examining the noise effects on speech, we choose one utterance in clean, and two noisy (train noise) conditions from AURORA-2J data corpus [8] with SNR of 20 dB and 10 dB. The MFCCs are calculated with 20 ms frame length and 10 ms frame shift. We then calculate the spectrum of the cepstral time series to investigate the noise effect on the modulation information of speech. For reverberant noise condition, we adopt the similar processing procedures. The reverberant speech is artificially generated by convolution between the clean speech (the same utterance as used for additive noise condition) with the impulse response of a smart-room (with RT of 650 ms) [7]. In both additive and reverberant noise conditions, the modulation spectrum is calculated using smoothed power spectrum of the cepstral time series. The modulation spectrum figures of the first order cepstral coefficient of clean and noisy speech are shown in Fig. 2. From Fig. 2a, it was found that, in noisy conditions, the low modulation components (less than 1 Hz) are enhanced, while the high modulation components (larger than 1 Hz) are attenuated. This modulation spectrum change shows that the MTF of the noise effect can be regarded as a low pass filtering to the cepstral temporal modulation, the lower the SNR, the smaller the end frequency of the filter. For reverberant condition, as shown in Fig. 2b (the two reverberant conditions correspond to the SMD with $d = 100$ cm and $d = 300$ cm), we can see that in reverberant condition, the MTF of the room acoustic has similar effect for the modulation spectrum as that in additive noise condition. But there are some differences in the changes in modulation components, for example, the low modulation components are enhanced (less than 4 Hz), while the high modulation components are attenuated (larger than 5 Hz). Either in additive noise or reverberant condition, the enhancement of the low modulation components while attenuating of the high modulation components means that the TMC of the time series is decreased.

3. PROPOSED TWO-STEP PROCESSING FOR TEMPORAL MODULATION NORMALIZATION

From section 2, we get the general conclusion of the noise effects on the temporal modulation spectrum of the cepstral coefficients, i.e., the TMC is decreased, and the high temporal modulation components are attenuated. Correspondingly, for reducing the noise effect, we should normalize these two effects for the noisy speech. We propose to use TMC normalization and edge-preserved smoothing algorithm to process the time series of cepstral coefficients thus to attenuate the noise effects.

3.1. Temporal modulation contrast (TMC) normalization

In some robust speech feature extraction methods for TMC enhancement, the dynamic range normalization is implicitly used. For example, the cepstral mean and variance normalization (CMVN) is often used. In our study, we find that we can explicitly use the dynamic range normalization processing to reduce the difference between clean speech and noisy speech either for additive noise or reverberant noise. However, for simplicity, in this study, we still adopt the CMVN as the dynamic range normalization processing. The CMVN is used as

$$\hat{c}\,(k,t) = \frac{c\,(k,t) - \bar{c}\,(k,t)}{\sigma\,(k)}, \qquad k = 0, 1, \cdots, \tag{1}$$

where $c(k,t)$ is the k-th cepstral coefficient with time index t. The cepstral mean $\bar{c}(k,t)$ and the standard deviation $\sigma(k)$ are calculated as follows.

$$\bar{c}\,(k,t) = \frac{1}{N}\sum_{t=1}^{N} c\,(k,t), \tag{2a}$$

$$\sigma\,(k) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(c\,(k,t) - \bar{c}\,(k,t)\right)^2}. \tag{2b}$$

In Eq. (2), N is the number of frames of the processed utterance. Traditionally, the CMVN is usually explained as to normalize the distribution of speech feature to be standard Gaussian distribution. In our MTF concept, it is explained as the TMC normalization which is used to equalize the modulation depth similar as used in image processing. After this CMVN processing, all the TMC of clean and noisy utterances are normalized to be in the same level, i.e., the TMC for noisy speech is enhanced as the same level of that of clean speech.

3.2. Edge-preserved smoothing on contrast normalized feature

In our study, we found that after the TMC normalization, there are some abrupt TMSs caused by the non-stationary noise, we need to smooth out those modulation components which are produced by the

noise artifacts while keeping speech modulation events unchanged. As suggested that most of the intelligibility information of speech is distributed in the low modulation frequency range (between 2 Hz and 16 Hz) [2, 5]. Following this idea, many temporal filtering methods are proposed. Most of them try to design a filter with modulation frequency in the range of speech modulation. Even some filters are estimated using data-driven methods. However, most of them adopt the average property of the speech modulation [5, 9]. For example, in the RASTA filtering, the filtering pass-band frequency range is around 0.2 Hz-16 Hz. It is better to adopt an adaptive filtering strategy to enhance speech modulation boundaries. Based on this consideration, we propose to use edge-preserved filtering to smooth the temporal trajectory while keeping the speech modulation boundary information [10]. The smoothing filter is designed as follows.

$$\tilde{c}(k,t) = \frac{\sum_{i=-m}^{m} w_k(t,i)\,\hat{c}(k,t-i)}{\sum_{i=-m}^{m} w_k(t,i)}, \qquad k = 0,1,2,\cdots \tag{3}$$

In Eq. (3), $w_k(t,i)$ is the weighting coefficient, m is the smoothing step order. The weighting coefficient is calculated as:

$$w_k(t,i) = w_{k_S}(t,i) \cdot w_{k_R}(t,i), \qquad k = 0,1,2,\cdots \tag{4}$$

$$w_{k_S}(t,i) = \exp\left\{-\frac{d_S^2(t,t-i)}{2\sigma_S^2}\right\} \tag{5a}$$

$$w_{k_R}(t,i) = \exp\left\{-\frac{d_R^2(\hat{c}(k,t),\hat{c}(k,t-i))}{2\sigma_R^2}\right\}. \tag{5b}$$

In Eq. (5), $d_S^2(t,t-i)$ is the Euclidian distance between time step t and $t-i$; the $d_R^2(\hat{c}(k,t),\hat{c}(k,k-i))$ is the Euclidian distance between the values of $\hat{c}(k,t)$ and $\hat{c}(k,t-i)$; the w_{k_S} formed as a Gaussian filter is used to smooth the time trajectory, while the w_{k_R} is used to keep the speech modulation regions by reducing the diffusion of different modulation events by considering the difference of the neighboring temporal values. By adjusting the controlling variances of the two parts σ_R^2 and σ_S^2, we can adaptively smooth the noise artifacts while keeping certain speech temporal modulation events (edges).

4. EVALUATIONS AND SPEECH RECOGNITION EXPERIMENTS

By the TMC normalization and edge-preserved smoothing, the temporal modulation of the cepstral coefficient of noisy speech is expected to be closer to that of the clean speech than the un-processed ones. The recognition performance based on the normalized cepstral coefficients should be improved.

4.1. The effects of the normalization processing on the temporal modulation structure

In order to show the effect of the temporal normalization processing, we apply this processing on the same speech utterance in clean and noisy conditions as used in section 2. For additive and reverberant noise conditions, the results for the first order cepstral coefficient are shown in Fig. 3. Comparing Figs. 2a and 2b with Figs. 3a and 3b, we can see that after the normalization processing, the TMSs of noisy speech are more close to those of clean speech, especially in the modulation frequency range between 2 Hz and 20 Hz.

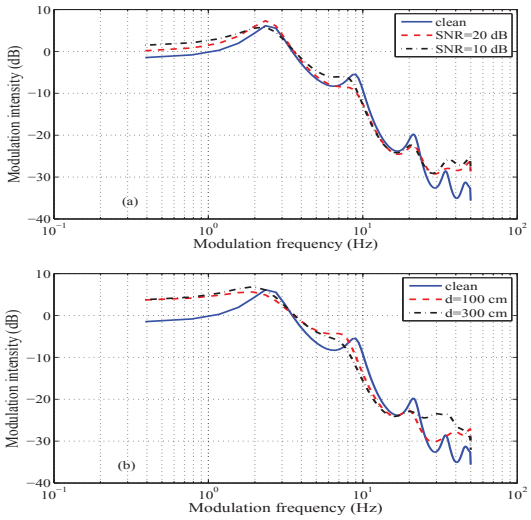

Fig. 3. Modulation spectrum after normalization processing for (a) additive noise and (b) reverberant noise conditions.

Table 1. Recognition rates and relative improvement (%)

Clean training	Set A	Set B	Set C	Overall	Relative
Baseline	46.78	48.21	49.44	47.91	—
AFE	79.43	77.96	75.88	78.13	***59.37***
Proposed	76.42	78.73	74.67	77.00	***57.26***
Multi-training	**Set A**	**Set B**	**Set C**	**Overall**	**Relative**
Baseline	88.66	79.96	82.67	83.98	—
AFE	93.21	88.88	90.63	90.96	***35.77***
Proposed	91.78	89.43	90.82	90.65	***33.52***

4.2. Speech recognition experiments

We tested the proposed normalization algorithm on speech recognition task in both additive noise and reverberant conditions in the following subsections.

4.2.1. Speech recognition in additive noise condition

The AURORA-2J data corpus was used in our experiments for additive noise condition test. The feature type and acoustic models used were the same as those used in the AURORA-2J experiments. For comparison, the ETSI advanced front-end (AFE) which is one of the best front-end processors was used [11]. The recognition results for clean and multi-conditional training are shown in Table 1. In Table 1, the recognition rates are the average for SNRs from 20 dB to 0 dB. The recognition rates in column with "Overall" means the average for testing sets A, B, and C. The column with "Relative" means relative improvement compared with baseline. From Table 1, we can see that the performance of the proposed algorithm can be comparable with that of the ETSI AFE algorithm for the additive noise condition.

4.2.2. Speech recognition in reverberant condition

For speech in reverberant condition, we used clean speech from the AURORA-2J database as the speech material. 8840 clean speech sentences were used to train the acoustic models. 1001 clean speech sentences convolved with the impulse responses of a smart room

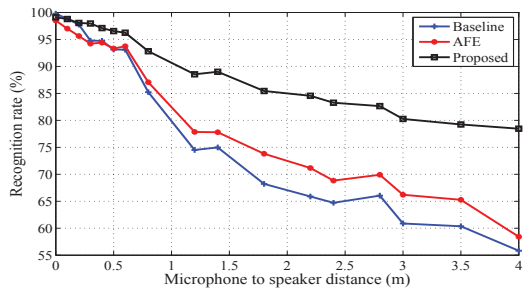

Fig. 4. Speech recognition under reverberant environment.

Table 2. Recognition rates (%) for noisy condition with both reverberant noise and additive noise

Method	Cond. 1	Cond. 2	Cond. 3	Cond. 4
Baseline	72.97	43.97	59.69	30.15
AFE	82.87	77.31	71.90	69.09
Proposed	***90.93***	***87.29***	***84.31***	***82.87***

were used as testing speech. The impulse responses depended on the speaker to microphone distances [7]. The feature type and acoustic models are configured the same as those used in AURORA-2J experiments. The recognition results are shown in Fig. 4. From Fig. 4, we can see that the proposed temporal modulation normalization algorithm significantly improve the robustness of speech recognition in the reverberant condition. Compared with the baseline performance, by averaging the recognition rates with SMD from 0.01 m to 4 m, our proposed algorithm has 51.28% relative improvement which is higher than that of ETSI AFE (10.17% relative improvement).

4.2.3. Speech recognition in noisy environment with both additive and reverberant noise

In real noisy environments, both reverberant and additive noise were presented. In order to examine the effectiveness of the proposed algorithm in those noisy environments, we simulated the real noisy environments by adding air conditioner noise to the reverberant speech in the smart room with regard to the SMD and SNR. Four testing conditions were simulated as: (Cond. 1) SMD=100 cm with SNR=20 dB, (Cond. 2) SMD=100 cm with SNR=10 dB, (Cond. 3) SMD=200 cm with SNR=20 dB and (Cond. 4) SMD=200 cm with SNR=10 dB. The feature types and acoustic models are the same as used in the subsections 4.2.1 and 4.2.2. The recognition results are shown in Table 2. From Table 2, we can see that, both the AFE and proposed algorithms improved the performance significantly in the real environment for the noisy speech recorded using the distance microphone. Especially, on average of the four testing conditions in Table 2, our proposed algorithm has relative improvement of 71.74% which outperforms the ETSI AFE algorithm with 48.86% relative improvement compared with the baseline performance.

5. CONCLUSION AND DISCUSSION

In this paper, we analyzed the noise effects on the TMSs of cepstral coefficients. Based on the analysis, we found that the noise not only changes the TMC (or dynamic range), but also attenuates high modulation components. Based on these findings, we proposed the TMC normalization (or dynamic range normalization) and edge-preserved smoothing to process the temporal modulation of cepstral coefficients. The purpose for the normalization is to normalize the TMS of noisy speech to those of clean speech. The proposed algorithm was tested on speech recognition in both additive and reverberant noise conditions. Compared with baseline algorithm, for additive noise, we got relative improvement of 57.25% for clean training condition, and 33.52% for multi-conditional training; for reverberant noise condition, we got 51.28% relative improvement; and for noisy condition with both reverberant and additive noise, we got 71.74% relative improvement (averaging on the four testing conditions in Table 2).

For more complex and adverse noise conditions, for example, in competitive speakers' speech, the MTF for the target speech is more complex. However, the sound sources possibly have different temporal modulation structures which are distributed in the temporal-frequency space. One source usually synchronizes its temporal modulation structure which is an important cue for separating interfere sound to solve this kind of cocktail party problem [4]. This will be our future work and final goal.

6. ACKNOWLEDGEMENTS

This study is supported by Knowledge Creating Communication Research Center.

7. REFERENCES

[1] P. C. Loizou. Speech enhancement: theory and practice, CRC Press, 2007.

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

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

[4] L. Atlas, and S. Shamma. "Joint Acoustic and Modulation Frequency," *EURASIP JASP*, No.7, 668-675, 2003.

[5] 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.

[6] T. Houtgast and H. J. M. Steeneken. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, 77 (3), 1069-1077, 1985.

[7] Neumann, J., Gasas, J. R., Macho, D., Hidalgo, J. R. "Integration of audio-visual sensors and technologies in a smart room," *Personal and Ubiquitous Computing*, Springer London, ISSN: 1617-4909 (print), 1617-4917 (online), 2007.

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

[9] X. Xiao, E. S. Chng, H. Li. "Temporal Structure Normalization of Speech Feature for Robust Speech Recognition," *IEEE Signal Processing Letters*, 14 (7), 500-503, 2007.

[10] M. Elad. "On the origin of the bilateral filter and ways to improve It," *IEEE Transactions On Image Processing*, 11 (10), 1141-1151, 2002.

[11] ETSI ES 202 050 V1.1.5. "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithms; compression algorithms," ETSI standard, 2007.