Modulation Spectrum Power-law Expansion for Robust Speech Recognition

Hao-Teng Fan, Zi-Hao Ye and Jeih-weih Hung
Department of Electrical Engineering, National Chi Nan University, Nantou, Taiwan
E-mail: s99323904@ncnu.edu.tw, s101323553@ncnu.edu.tw, jwhung@ncnu.edu.tw

***Abstract*—In this paper, we present a novel approach to enhancing the speech features in the modulation spectrum for better recognition performance in noise-corrupted environments. In the presented approach, termed modulation spectrum power-law expansion (MSPLE), the speech feature temporal stream is first pre-processed by some statistics compensation technique, such as mean and variance normalization (MVN), cepstral gain normalization (CGN) and MVN plus ARMA filtering (MVA), and then the magnitude part of the modulation spectrum (Fourier transform) for the feature stream is raised to a power (exponentiated). We find that MSPLE can highlight the speech components and reduce the noise distortion existing in the statistics-compensated speech features. With the Aurora-2 digit database task, experimental results reveal that the above process can consistently achieve very promising recognition accuracy under a wide range of noise-corrupted environments. MSPLE operated on MVN-preprocessed features brings about 55% in error rate reduction relative to the MFCC baseline and significantly outperforms the single MVN. Furthermore, performing MSPLE on the lower sub-band modulation spectra gives the results very close to those from the full-band modulation spectra updated by MSPLE, indicating that a less-complicated MSPLE suffices to produce noise-robust speech features.**

I. INTRODUCTION

Broadly speaking, the state-of-the-art automatic speech recognition (ASR) system can perform well in a well-controlled laboratory environment, while its performance usually degrades in real-world applications. The environmental mismatch that causes the above performance degradation is from the various interfering sources such as noise/interference and channel distortion/fading. To solve or conquer the problem, a great number of noise robustness methods have been proposed in different stages of speech recognition process. Initially, the statistics normalization is operated on the *temporal* domain of speech features, and the respective methods include cepstral mean normalization (CMN) [1], mean and variance normalization (MVN) [2], cepstral histogram normalization (CHN) [3] and MVN plus ARMA filtering (MVA) [4]. Later, the concept of statistics normalization is further used to process the *modulation spectral* domain of speech features, and the methods of spectral histogram equalization (SHE) [5], magnitude ratio equalization (MRE) [5] and sub-band statistics normalization techniques [6] are accordingly developed. By and large, the paring of temporal- and modulation spectral-domain methods can give superior performance relative to the component single domain method.

In our recent research, we proposed three new modulation spectral-processing methods: modulation spectrum replacement (MSR) [7], modulation spectrum filtering (MSF) [7] and modulation spectrum exponential weighting (MSEW) [8]. These three methods attempt to reduce the mismatch between the clean and noise-corrupted speech in modulation spectrum and make the updated speech features more noise-robust and thereby producing superior recognition performance. Briefly speaking, MSR and MSF apply a uniform reference magnitude spectrum to replace/scale the modulation spectrum of different feature streams at different signal-to-noise ratios (SNRs), while MSEW further refines MSR and MSF by adjusting the proportion of the underlying reference magnitude spectrum according to the noise level of the environment. It has been shown that these three methods can enhance the MVN-preprocessed cepstral features in noise robustness and give rise to the improvement of recognition accuracy relative to MVN alone.

A common characteristic of MSR, MSF and MSEW is that they require a reference magnitude spectrum, which is usually estimated by the features of the clean training set. However, the clean training data are not always available or the corresponding quantity is not large enough to produce a good estimate of the reference magnitude spectrum. In addition, the reference magnitude spectrum varies with the used feature type and the pre-processing method for the original features. Adopting a new feature type and/or a pre-processing method when using any of MSR, MSF and MSEW requires the re-estimation of the reference magnitude spectrum.

In light of the aforementioned observations, we present a novel modulation spectrum updating algorithm in this paper. In this novel algorithm, termed modulation spectrum power-law expansion, with MSPLE as the shorthand notation, the magnitude portion of the modulation spectrum for each feature time-series is raised to a power larger than one. As a result, any (magnitude) spectral component originally greater/less than one is to be further amplified/reduced by MSPLE. Due to the fact that the feature time series in general reveals a low-pass characteristic, the operation of MSPLE will further highlight its low-pass portion, which often corresponds to more important information for speech recognition [9]. In comparison with MSR, MSF and MSEW, the new method MSPLE does not need any reference magnitude spectrum and thus can be done more flexibly with less computation complexity. Experiments conducted on the Aurora-2 database [10] shows that MSPLE can improve the MFCC features pre-processed by some specific mean-

removing process (such as MVN, CHN and MVA) in recognition accuracy, and MSPLE performs almost equally well as MSR, MSF and MSEW.

The remainder of this paper is organized as follows: Section II describes the procedures of the presented MSPLE together with its several properties. The experimental setup is given in Section III, and Section IV contains a series of experimental results and the corresponding analysis and discussions. Finally, a brief concluding remark is provided in Section V.

II. PROPOSED METHOD

Consider using the zero-mean MFCCs as the baseline features for speech recognition, and let $\{x_m[n], 0 \le n \le N-1\}$ denotes the m^{th} MFCC time series of an arbitrary sequence in the training and testing sets. We attempt to update this feature time series $\{x_m[n]\}$ via its Fourier transform, i.e., the modulation spectrum, hoping that the resulting new feature time series, denoted by $\{\tilde{x}_m[n], 0 \le n \le N-1\}$, can be more noise-robust. For the purpose of a compact presentation, we omit the cepstral index m in later discussions, unless otherwise specified.

The presented algorithm, modulation spectrum power-law expansion (MSPLE) consists of the following three steps. First, we convert the time series $\{x[n], 0 \le n \le N-1\}$ into the modulation spectrum via N-point DFT:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi nk}{N}}, \qquad k = 0, 1, \dots, N-1, \qquad (1)$$

and then use $A[k]$ and $\theta[k]$ to represent the magnitude and phase parts of $X[k]$, respectively. Note that the sequence $\{X[k], 0 \le k \le N-1\}$ is conjugate symmetric, and the component corresponding to the highest frequency is $X[\lceil N/2 \rceil]$, where $\lceil . \rceil$ denotes the ceil operation.

Next, the magnitude part, $A[k]$, is raised to power α:

$$\tilde{A}[k] = (A[k])^{\alpha}, \qquad k = 0, 1, \dots, N-1, \qquad (2)$$

where $\tilde{A}[k]$ denotes the new magnitude part. As shown in Eq. (2), the exponentiation operation is carried out on the all frequency components, and thus here the MSPLE is implemented in a "full-band" manner.

Finally, the new time series $\{\tilde{x}[n]\}$ is obtained by the N-point inverse DFT of the combination of the new magnitude part $\tilde{A}[k]$ and the original phase part $\theta[k]$:

$$\tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} (\tilde{A}[k] e^{j\theta[k]}) e^{-j\frac{2\pi nk}{N}}, \quad k = 0, 1, \dots, N-1 \qquad (3)$$

Apparently, the exponentiation operation shown in Eq. (2) with the power $\alpha > 1$ will enlarge the dynamic range of the magnitude spectrum $A[k]$ (as long as the maximum value of $A[k]$ is greater than 1).

Figure 1 shows the power spectral density (PSD) curves (a smoothed version of the magnitude spectra) of the MVN-preprocessed MFCC $c1$ for an utterance "MSA_ZZZ73A.08" in the Aurora-2 database at three signal-to-noise ratios (SNRs): clean, 20 dB and 10 dB. From this figure, we first find the modulation spectra exhibit a low-pass characteristic (except for the near-DC portion, which is significantly diminished by the MVN process). Second, the noise distortions left behind by MVN are primarily located at the relatively high frequency

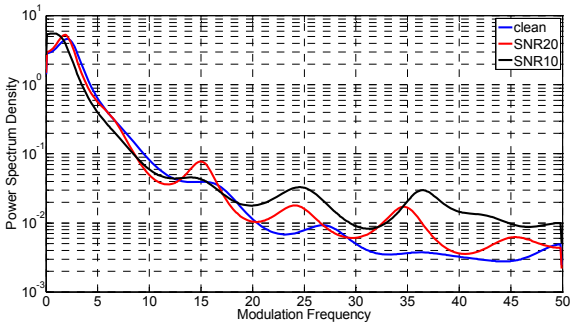

Figure 1. The MVN-processed MFCC c1 PSD curves of an utterance at three SNR cases: clean, 20 dB and 10 dB.

region. By virtue of the presented MSPLE process, the low-pass characteristic of the PSD shown in Figure 1 can be further emphasized, implying that MSPLE highlights the lower frequency components that are commonly viewed to be more beneficial for the speech recognition more than higher frequency component. Furthermore, the noise distortions dwelled at the high frequency region shown in Figure 1 can be reduced by MSPLE since the corresponding values are less than 1. As mentioned earlier, the lower modulation frequency components are more important among the entire modulation spectrum. Thus here we attempt to modify the original MSPLE in the way that the exponentiation operation is just carried out *at the lower modulation frequencies* as follows:

$$\tilde{A}[k] = (A[k])^{\alpha}, \ k = 0, 1, \dots, M, N-M, \dots, N-1, \qquad (4)$$

where M stands for the upper cutoff frequency index of the selected low-band, and $M \le \lceil N/2 \rceil$. Therefore, the presented MSPLE using Eq. (4) is further termed as "low-band MSPLE", and the ratio of low-band to full-band in bandwidth,

$$r = M/\lceil N/2 \rceil, \qquad (5)$$

determines the proportion of frequency components to be processed by MSPLE. Obviously, decreasing the value of r in Eq. (5) results in lower computation but probably a less effective performance of MSPLE.

III. EXPERIMENTAL SETUP

We evaluated the presented MSPLE with the speech recognition task of the Aurora-2 database [10], which consists of connected English digit strings. For the recognition environment, there are three test sets: The utterances of Test Sets A and B are affected by either of eight types of additive noise (subway, babble, etc.), and those in Test C is by two types of additive noise and a channel distortion relative to the clean training set. The SNR level of each testing utterance ranges from 20 and -5 dB, with an interval of 5 dB. Each utterance in the clean training set and three noise-corrupted testing sets is converted into a sequence of 13-dimensional mel-frequency cepstral coefficients (MFCC, c0-c12). The frame length and frame shift are set to 25 ms and 10 ms, respectively, so the MFCC sequence is within the modulation frequency band [0, 50 Hz]. The 13-dim static MFCCs plus their first and second order derivatives are then the components of the 39-dimensional feature vector used as the

Table 1. The optimal exponent α in MSPLE with respect to different pre-processing methods derived from the recognition results of the development set.

Preprocessing method	The optimal power-law value α
MFCC baseline	0.6
MVN	1.8
CGN	1.6
MVA	1.6

Table 2. Recognition accuracy (%) achieved by various methods for the Aurora-2 clean condition training task averaged across the SNRs between 0 and 20 dB. RR (%) is the relative error rate reduction over the MFCC baseline.

Method	Set A	Set B	Set C	Avg.	RR
MFCC (baseline)	59.24	56.37	67.53	59.75	—
MFCC+MSPLE	67.52	71.29	67.25	68.97	22.91
MVN	73.83	75.02	75.08	74.55	36.77
MVN+MSPLE	81.79	82.08	80.79	81.71	54.56
CGN	79.63	80.95	79.72	80.18	50.76
CGN+MSPLE	83.05	83.88	82.63	83.30	58.51
MVA	78.15	79.17	79.12	78.75	47.20
MVA+MSPLE	82.36	82.55	82.31	82.43	56.35
MVN+MSR	80.71	82.32	79.73	81.16	53.19
MVN+MSF	82.22	82.95	82.82	82.63	56.84
MVN+MSEW	82.58	83.24	82.82	82.89	57.49

Table 3. Recognition accuracy (%) achieved by various pre-processing methods and the low-band MSPLE with the various bandwidth ratio r defined in Eq. (5).

Method		Set A	Set B	Set C	Avg.
MVN+MSPLE	$r = 1$	81.79	82.08	80.79	81.71
	$r = 1/2$	81.80	82.10	80.94	81.75
	$r = 1/4$	81.69	81.80	80.86	81.57
	$r = 1/8$	77.81	78.14	77.44	77.87
CGN+MSPLE	$r = 1$	83.05	83.88	82.63	83.30
	$r = 1/2$	83.03	83.85	82.67	83.28
	$r = 1/4$	82.96	83.85	82.58	83.24
	$r = 1/8$	82.35	83.01	81.98	82.54
MVA+MSPLE	$r = 1$	82.36	82.55	82.31	82.43
	$r = 1/2$	82.34	82.54	82.31	82.41
	$r = 1/4$	82.27	82.38	82.29	82.32
	$r = 1/8$	79.72	79.88	79.67	79.77

baseline features. On the other hand, each of the robustness approaches to be evaluated is performed on the 13-dim static MFCCs only, and then the 26-dim dynamic MFCCs are computed accordingly. With the feature vectors in the clean training set, the HMMs for each digit and silence are trained. Each digit is modeled by an HMM with 16 states, left-to-right, with three Gaussian mixtures for each state [11].

According to (5), the key component in the presented MSPLE algorithm is the power-law factor α. In order to obtain a good selection for this parameter, we use the 8440 utterances (with five SNRs: clean, 20 dB, 15 dB, 10 dB and 5 dB) in the training set for the mode of multi-condition training of the Aurora-2 database as the development set. For the baseline MFCC features and the MFCC features pre-processed by either of MVN, CGN and MVA, the factor α is respectively set to 0.0, 0.2, ..., 2.0, with an interval of 0.2, in MSPLE processing. The power factors that result in the optimal recognition accuracy for different types of features in the development set are then selected for MSPLE processing in the testing set, and they are shown in Table 1.

From this table, we see that in MSPLE the power factor α that behaves nearly optimal in the development set is always greater than 1.0 for those cases when the MFCC features is pre-processed by MVN, CGN or MVA. These results partially support our previous statements that expanding the scale of the modulation spectra helps to enlarge the difference between the low-frequency and high-frequency portions and thereby the features are more noise-robust. As for the plain MFCC, however, using the power factor α less than 1.0 in MSPLE (implying the lower-frequency modulation spectrum shrinks) gives rise to better results probably due to the reduction of the distortions primarily located in the low frequencies.

IV. EXPERIMENTAL RESULTS

In the first part of experiments, we compare the presented MSPLE with several noise-robustness methods in recognition accuracy. Table 2 shows the recognition accuracy rates for different features processed by any of various methods, including the presented MSPLE which sets the power factor according to Table 1. From this table, we have the following observations:

1. The three well-known temporal processing methods, MVN, CGN and MVA, benefit the MFCC features a lot in prompting the recognition performance for all three Test Sets. CGN behaves the best, followed by MVA and then MVN. The presented MSPLE (with the power factor α being 0.6) outperforms the MFCC baseline for Test Sets A and B, but degrades the recognition accuracy slightly for Test Set C. Therefore, MSPLE seems not quite appropriate to process the plain MFCC directly.
2. In contrast with the case of plain MFCC, MSPLE performs very well for those features pre-processed by any of MVN, CGN and MVA. For example, MSPLE gives rise to 7.16%, 3.12% and 3.68% in absolute accuracy improvement for MVN-, CGN- and MVA-processed features, respectively. These results indicate that the presented MSPLE are well additive to these pre-processing methods and can produce further noise-robust features.
3. With MVN as the preprocessing method, the new approach MSPLE behaves better than MSR and worse than MSF and MSEW. As mentioned earlier, MSR, MSF and MSEW rely on a reference magnitude spectrum from the clean training data, which has to be given *a priori*. On the contrary, MSPLE updates the modulation spectrum in a *blind* manner without any information of the clean data. Therefore, it comes as no surprise that MSPLE is probably less effective among the four approaches discussed here.

In the second part, the low-band MSPLE given in Eq. (4) with three different assignments of the bandwidth ratio (the parameter r defined in Eq. (5)), 50%, 25% and 12.5%, is evaluated, and the corresponding recognition performance is shown in Table 3. For simplicity, in this table we just list the results of MSPLE for the features pre-processed by any of

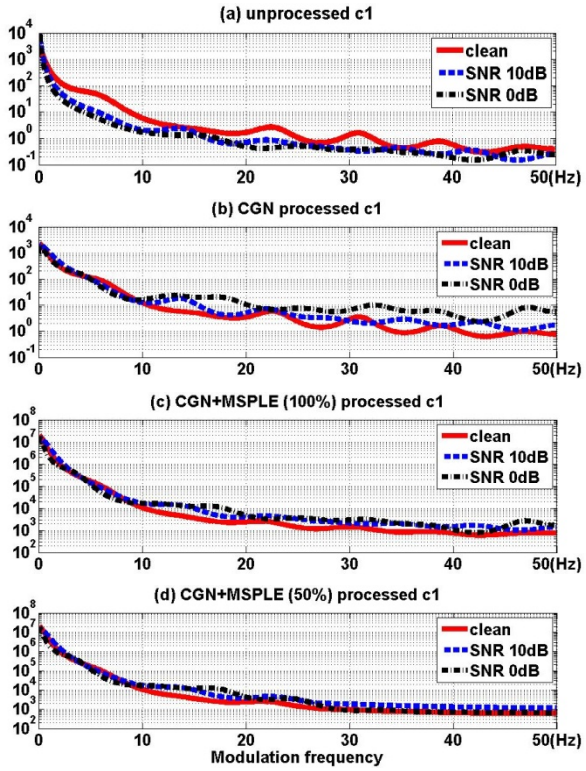

Figure 2. The MFCC c1 PSD curves processed by various compensation methods: (a) the MFCC baseline (no compensation), (b) CGN, (c) CGN plus full-band MSPLE, and (d) CGN plus low-band MSPLE ($r = 50\%$)

MVN, CGN and MVA. From Table 3, we find that the low-band MSPLE can behave as well as full-band MSPLE even when the width of the processed low-band is just 25% of the full-band (approximately within the band [0, 12.5 Hz]), and thus we can reduce the computation complexity of MSPLE by as much as 75% without the expense of degrading the recognition performance. However, when the processed low-band is further reduced to one-eighth (12.5%) of the full-band, the corresponding MSPLE performs worse than the full-band counterpart, while it can still enhance the original features in recognition accuracy. For example, the method "CGN+MSPLE (1/8)" gives the averaged accuracy of 81.98%, higher than 80.18% achieved by CGN alone.

Finally, we examine the ability of MSPLE to reduce the modulation spectrum distortion caused by noise. Figures 2(a)-2(d) plot the PSD curves of the first MFCC $c1$ for an utterance "MAH_27O6571A" in the Aurora-2 database for three SNR levels, clean, 10 dB and 0 dB (with airport noise) before and after various processes (CGN, CGN plus MSPLE with $r = 100\%$, and CGN plus MSPLE with $r = 50\%$), respectively. First, Fig. 2(a) shows that the noise results in a significant PSD mismatch over the entire modulation frequency band [0, 50 Hz] of the original MFCC $c1$ sequence. Second, by comparing Fig. 2(b) with Fig. 2(a) we find that CGN primarily diminishes the PSD distortion in the lower frequency band (around [0, 10 Hz]), and the PSD mismatch still remains at the higher frequencies. However, CGN can provide significant accuracy improvement relative to MFCC baseline, as evident in Table 2, revealing the fact that taking care of low-frequency distortion suffices to promote the noise robustness significantly. Finally, Figs. 2(c) and 2(d) show that the presented MSPLE can further reduce the PSD distortion less treated by CGN, especially for the frequency range [10 Hz, 20 Hz]. In particular, there is no substantial difference between the PSD curves obtained by full-band MSPLE and low-band MSPLE ($r = 50\%$) since the high frequency components are relatively small, which partially explains why these two methods behave very similar in recognition performance as shown in Table 3.

V. CONCLUSIONS

The presented novel noise-robustness approach, modulation spectrum power-law expansion (MSPLE), enhances the speech features in noise robustness by expanding the dynamic range of the modulation spectrum. In comparison with our previous proposed modulation spectrum processing methods, MSPLE can be implemented in a simpler manner while obtain similar performance. In the future work, we will pursue an adaptive way to tune the used exponent factor in MSPLE according to the signal-to-noise ratio (SNR) of the processed utterance in order to achieve superior noise-robust performance.

REFERENCES

[1] S. Tibrewala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition," in *Proceedings of the Eurospeech Conference on Speech Communications and Technology*, 1997.

[2] S. Yoshizawa et al., "Cepstral gain normalization for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209-212, 2004.

[3] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3), pp. 845-854, 2006.

[4] C. P. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), pp. 257-270, 2007.

[5] L. C. Sun and L. S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transaction on Audio, Speech and Language Processing*, 20(3), pp. 828-843, 2012.

[6] W. H. Tu, S. Y. Huang and J. W. Hung, "Sub-band modulation spectrum compensation for robust speech recognition," in *Proceedings of the Automatic Speech Recognition and Understanding*, pp. 261-265, 2009.

[7] J. W. Hung, W. H. Tu and C. C. Lai, "Improved modulation spectrum enhancement methods for robust speech recognition," *Signal Processing*, 92(11), pp. 2791-2814, 2012.

[8] H. T. Fan, Y.-C. Lian and J. W. Hung, "Modulation spectrum exponential weighting for robust speech recognition," in

Proceedings of the International Conference on ITS Telecommunications, pp. 812-816, 2012

[9] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proceedings of the Eurospeech Conference on Speech Communications and Technology*, 1997.

[10] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181-188, 2000.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge, UK, 2006.