MODULATION SPECTRUM EQUALIZATION FOR ROBUST SPEECH RECOGNITION

Liang-che Sun , Chang-wen Hsu, and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
Taiwan, Republic of China
lgsun@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Two approaches for modulation spectrum equalization are proposed for robust feature extraction in speech recognition. In both cases the temporal trajectories of the feature parameters are first transformed into the modulation spectrum. In the spectral histogram equalization (SHE) approach, we equalize the histogram of the modulation spectrum for each utterance to a reference histogram obtained from clean training data. In the magnitude ratio equalization (MRE) approach, we equalize the magnitude ratio of lower to higher frequency components on the modulation spectrum to a reference value also obtained from clean training data. Preliminary experimental results performed on the AURORA 2 testing environment indicate that significant performance improvements are achievable with these approaches, when integrated with cepstral mean and variance normalization (CMVN), for all testing sets A, B, and C, all types of noise, for all SNR values. We also show that the approach of magnitude ratio equalization (MRE) offers additional performance improvements when integrated with other more advanced feature normalization approaches such as histogram equalization (HEQ) and higher-order cepstral moment normalization (HOCMN).

Index Terms— Modulation spectrum, feature normalization, robust feature extraction, temporal filter

1. INTRODUCTION

The performance of speech recognition systems is very often degraded due to the mismatch between the acoustic conditions of the training and testing environments. A very popular approach for handling this problem is to try to normalize the statistical behavior of the speech features in order to reduce the effect of such mismatch under various environmental conditions. Cepstral mean subtraction (CMS) [1], cepstral mean and variance normalization (CMVN) [2], histogram equalization (HEQ) [3], and higher-order cepstral moment normalization (HOCMN) [4] are typical examples of such techniques. CMS and CMVN normalize the first-order and/or the second-order feature moments, and HOCMN further normalizes other moments of higher orders. HEQ, on the other hand, equalizes the histogram of speech features to some reference cumulative distribution function (CDF). In general, these techniques all seek to normalize the distributions of the speech features.

Another approach to reducing the above mismatch is to try to filter the time trajectories of the speech features, or to perform filtering in the modulation spectrum. RASTA filtering [5] [6] or other similar approaches with filters designed by data-driven methods based on different criteria such as linear discriminant analysis (LDA) [7], principle component analysis (PCA) [8], and minimum classification error (MCE) [10] are good examples of this approach. Properly using information induced from the modulation spectrum [12] or performing square-root Wiener filtering on the modulation spectrum [13] are also good examples. It has been shown in most studies that the modulation spectrum around 4 Hz is the most useful for speech recognition [5][9][10][11].

In this paper, we propose a new approach for modulation spectrum equalization in which the modulation spectra of noisy speech utterances are equalized to those of clean speech. This includes two equalization techniques. The first is to equalize the cumulative density functions (CDFs) of the modulation spectra of clean and noisy speech, such that the differences between them are reduced. The second is to equalize the magnitude ratio of lower to higher components in the modulation spectrum, which also reduces the difference between the modulation spectra of clean and noisy speech. Experiments performed on the AURORA 2 testing environment offered very encouraging results. The rest of the paper is organized as follows. In section 2, the proposed approach is presented. In sections 3 and 4 the experimental setup and results are reported. Concluding remarks are finally presented in section 5.

2. PROPOSED APPROACH

As shown in Figure 1, given a sequence of feature vectors $\{x(n),n=1,2,...,N\}$ for an utterance, each including D feature parameters, where n is the time index, and $d=1,...,D$ is the parameter index,

$$x(n)= [x(n,1),x(n,2), ...,x(n,d),...,x(n,D)]^T , n=1, ...,N. \quad (1)$$

Then the time trajectory of the d-th parameter of $\{x(n),n=1,2,...N\}$ is the sequence $[x(1,d)\ x(2,d)\ ...x(N,d)]$, denoted as $y_d(n)$, where $y_d(n)= x(n,d)$. Now we can transform the temporal sequence $y_d(n)$ to the modulation

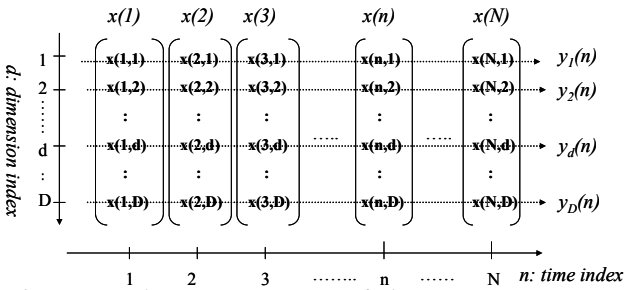


Figure 1: *The representation of the time trajectories of feature parameter sequences*

spectrum [10],

$$Y_d(k) = \sum_{n=0}^{N-1} y_d(n) \cdot \exp(-j2\pi nk / N), \quad (2)$$
$$k = 0,1,2,..,N-1; \quad d = 1,2,....,D,$$

where k is the frequency index of the discrete Fourier transform. The two techniques proposed here can then be performed with $Y_d(k)$. If complexity is a concern for real-time applications, a fixed value N optimized for the FFT algorithm can be chosen and $Y_d(k)$ can be obtained window-by-window.

2.1. Spectral Histogram Equalization (SHE)

Histogram equalization (HEQ) has been shown to be very useful in image processing and speech feature normalization. Here, we try to borrow this concept but apply it in the modulation spectrum, or $Y_d(k)$ as in equation (2). This is referred to as spectral histogram equalization (SHE). Let $Y_{d,test}(k)$ represent the modulation spectrum of a testing utterance. Such a modulation spectrum for the MFCC parameter c0 for a typical example test utterance of AURORA 2 is shown in Figure 2. We can observe that this modulation spectrum is greatly altered when the SNR is degraded to 10 dB or 0 dB. In general $Y_d(k)$ is a complex number, but here we only consider equalizing the magnitude $|Y_d(k)|$, while keeping the phase unchanged. We first calculate the cumulative distribution function (CDF) of the magnitudes of the modulation spectra, $|Y_d(k)|$, for all utterances in the clean training data of AURORA 2 to be used as the reference CDF, $\mathrm{CDF}_{\mathrm{ref}}[\cdot]$. For any test utterance, the CDF for its modulation spectrum magnitude, $|Y_{d,test}(k)|$, can be similarly obtained as $\mathrm{CDF}_{\mathrm{test}}[\cdot]$. Hence the equalized magnitude of modulation spectrum $|\hat{Y}_{d,test}(k)|$ is

$$\left|\hat{Y}_{d,test}(k)\right| = CDF_{ref}^{-1}(CDF_{test}[\left|Y_{d,test}(k)\right|]) \quad , \quad (3)$$

where $\mathrm{CDF}^{-1}{}_{\mathrm{ref}}[\cdot]$ is the inverse of the cumulative distribution function. This is the spectral histogram equalization (SHE), and after this process the statistical distribution of $|Y_{d,test}(k)|$ is better matched to that of the clean training speech data.

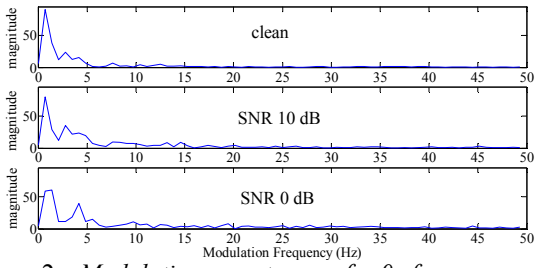


Figure 2: *Modulation spectrum of c0 for an example utterance in the AURORA 2 corpus under clean, 10 dB, and 0 dB SNR conditions, where the horizontal scale is the modulation frequency (Hz), and the vertical scale is the magnitude.*

2.2. Magnitude Ratio Equalization (MRE)

For a speech utterance, we first define a magnitude ratio (MR) for lower to higher frequency components for each parameter index d as follows:

$$MR_d = \frac{\sum_{k=0}^{k_C} |Y_d(k)|}{\sum_{k=k_C+1}^{[\frac{N}{2}]+1} |Y_d(k)|} \quad , \quad (4)$$

where k_c is the cut-off frequency used here, N is the order of the discrete Fourier transform, and $[N/2]$ is a function that returns the largest integer less than or equal to $N/2$. Thus MR_d is simply the ratio of sum of the lower frequency components to that of the higher frequency components on the modulation spectrum, where the lower and higher frequency components are divided by k_c. It is well known that for the modulation spectrum of speech signals the major signal components are in the lower frequencies, and those in the higher frequencies are primarily non-speech, or noise. Therefore MR_d as obtained in equation (4) can be seen as an indicator for the noise conditions of a given utterance.

The distribution of the value of MR_d for k_c taken as 5 Hz (the selection of this value will be discussed later on) for the parameter c0 for all test utterances in AURORA 2 including all testing sets at different SNR values is shown in Figure 3. We can observe from this figure that the mean value of MR_d is degraded when SNR is degraded, and thus MR_d is highly correlated with SNR. It is therefore reasonable to equalize the value of MR_d for a noisy utterance to a reference MR_d value obtained from clean training data.

We first calculate the average of MR_d for all utterances in the clean training data of AURORA 2 as the reference value $\mathrm{MR}_{d,\mathrm{ref}}$. Likewise, we then calculate the value of MR_{d} for each test utterance as $\mathrm{MR}_{d,\mathrm{test}}$. We thus define a scaling factor F_d as

$$F_d = \frac{MR_{d,ref}}{MR_{d,test}} \quad . \quad (5)$$

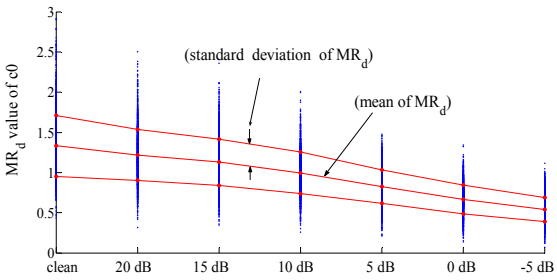


Figure 3: *The distribution of the magnitude ratio (*MR_d*) values of c0 for all testing utterances in AURORA 2 for all sets at all SNRs. Each point represents the* MR_d *value of c0 for an utterance.*

With F_d in equation (5), we then equalize the magnitude of the modulation spectrum for the test utterance $|Y_{d,test}(k)|$ as

$$\left|\hat{Y}_{d,test}(k)\right| = \begin{cases} F_d^{\,p} \cdot \left|Y_{d,test}(k)\right| & , k \le k_C \\ \dfrac{1}{F_d^{\,(1-p)}} \cdot \left|Y_{d,test}(k)\right| & , k > k_C \end{cases} \quad (6)$$

where $0<p<1$ is the weighted-power for the scaling factor. For example, if p=0.5, we use the same scaling factor to enhance the lower frequency components (or speech) and also to suppress the higher frequency components (or noise) in order to make the values of $MR_{d,test}$ identical to those of $MR_{d,ref}$. If p=0.3, we still make the values of $MR_{d,test}$ identical to $MR_{d,ref}$, but the lower frequency components are less enhanced while the higher ones are more suppressed. It will be shown later that the best values of k_c and p can be determined empirically.

2.3. The Overall Framework of the Proposed Approach

The overall framework of the approach proposed here is shown in Figure 4. We first perform feature normalization (CMVN, HEQ, or HOCMN) on both the training data and each test utterance before transforming them to the modulation spectrum. After then performing SHE and MRE, each test utterance has its own modulation spectrum histogram $CDF_{ref}[.]$ and $MR_{d,test}$ values, and thus is transformed individually. This is different from many conventional temporal filtering approaches, in which the same transformation (or set of filter coefficients) is used for different utterances and different noise conditions.

3. EXPERIMENTAL SETUP

The above approaches were evaluated under the AURORA 2 testing environment with an English connected-digit string corpus. Two training conditions (clean-condition and multi-condition) and three testing sets (sets A, B, and C) are defined in AURORA 2 [14]. In clean-condition training the acoustic models are trained on clean speech only, while in multi-condition training the models are trained using a

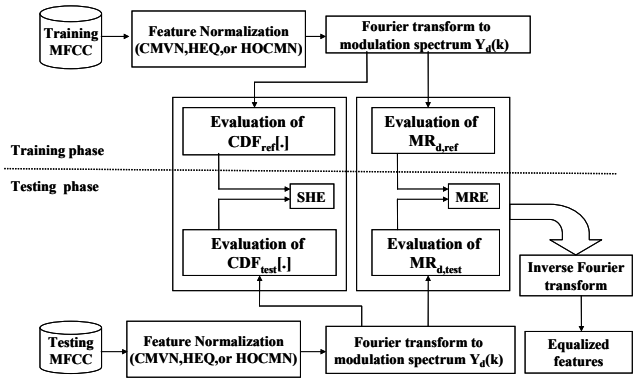


Figure 4: *The overall framework of modulation spectrum equalization techniques.*

corpus with both clean and noisy speech. The testing set A includes four different types of noise used in multi-condition training (subway, babble, car and exhibition), while the testing set B includes another four different types of noise not used in multi-condition training (restaurant, street, airport and train station). The testing set C then includes two noise types respectively from sets A and B (subway and street), plus additional convolutional noise. Five different SNR values ranging from 20 dB to 0 dB were tested in each case. Whole-word HMM models were used as specified by AURORA 2. Each word had 16 states and 3 Gaussian mixtures per state. The speech features were extracted by the AURORA WI007 front-end, which converted each signal frame into 13 cepstral coefficients (MFCCs, c0-c12), on which all the modulation spectrum equalization techniques proposed above were performed. The first and second derivatives were then computed from the equalized cepstral coefficients and used as well in the tests. The implementation of all the approaches tested here was based on the entire utterance; that is, N in equation (2) is the number of frames in an utterance.

4. EXPERIMENTAL RESULTS

4.1. Selection of Cut-Off Frequency k_c and Weighted-Power p in MRE

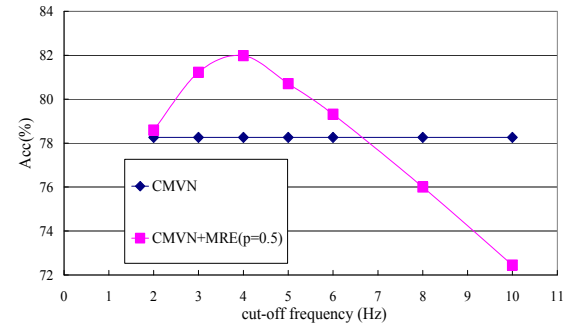
As can be found in equation (6) of section 2.2, the cut-off frequency k_c and weighted-power p play the key role in the proposed MRE technique. Since the cut-off frequency is very possibly the most important parameter here, we first set p=0.5 when choosing the optimal value of k_c. In Figure 5 (a), we show the word accuracy averaged over all noise types and all SNR values in sets A, B, and C for CMVN followed by MRE alone (without performing SHE), given the cut-off frequency k_c, where k_c is shown in Hz. From this figure we can see that a 4 Hz cut-off frequency is the best choice for MRE when p is 0.5. This makes good sense since the syllabic-rate in the AURORA 2 testing corpus is about 4

Hz, so if we choose this as the cut-off frequency, we will have the primary parts of speech information in the lower frequency band, in turn making MR_d in equation (4) a good candidate for equalization. This is consistent with earlier findings [5][9][10][11]. Also it is clear in Figure 5 (a) that with MRE with k_c set to 4 Hz, significant improvements can be obtained over CMVN alone.

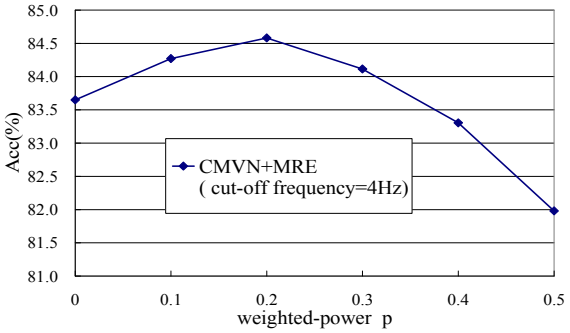
We then investigated optimal values for weighted-power p, assuming k_c is set to 4 Hz. In Figure 5 (b), word accuracy is shown as in Figure 5 (a), except for different weighted-power p with k_c set to 4 Hz. It is clear from Figure 5 (b) that p=0.2 gives the best results for MRE, and significant improvements can again be obtained as compared to p=0.5. Since the scaling factor F_d is usually larger than 1, this result (p=0.2<0.5) implies that increasing the suppression of higher frequency components (or noise) brings more benefits than enhancing lower frequency components (or speech) when we keep the value of $MR_{d,\text{test}}$ identical to that of $MR_{d,\text{ref}}$. This is also reasonable.

4.2. Performance of Modulation Spectrum Equalization Integrated with CMVN

The initial experimental results for clean condition training are shown in Table 1 for average results over all cases in testing sets A, B, and C and the overall average. The first two rows (1) (2) are for the MFCC baseline (using c0 instead of log energy) and CMVN respectively, and serve as the baselines for comparison. The last column of Table 1 is the error rate reduction with respect to CMVN (row (2)). Row (3) is CMVN followed by RASTA filtering, providing a limited error rate reduction of 0.45% over CMVN. Rows (4) and (5) are CMVN followed by PCA-derived (with filter length L=15) [8] and LDA-derived temporal filtering [7][8] (with filter length L=5), offering error rate reductions of 12.56% and 18.35% over CMVN. Rows (6), (7), (8) are then respectively the results of CMVN followed by SHE alone, MRE alone, and SHE+MRE, all with the best cut-off frequency k_c of 4 Hz and the best weighted-power p of 0.2 as selected in Section 4.1 above. Clearly, CMVN followed by either SHE or MRE in rows (6) and (7) provides significantly better results than CMVN alone (row (2)), or followed by temporal filtering approaches such as RASTA(row (3)), PCA (row (4)) or LDA (row (5)) under all test conditions. Also, MRE (row (7)) is able to provide much better results than SHE (row (6)); using both techniques (row (8)) is more effective than either single technique. The approaches of modulation spectrum equalization proposed in this paper --- SHE alone (row (6)), MRE alone (row (7)) and their integration (row (8)), all following CMVN --- are able to offer 23.64%, 29.07%, 29.39% error rate reduction with respect to CMVN alone. Moreover, MRE alone (row (7)) is better than SHE alone (row (6)) in all cases, and the combination of SHE and MRE (row (8)) is only slightly better than MRE alone (row(7)). Hence in many cases we may use MRE alone for simplicity.



(a)



(b)

Figure 5: *Recognition accuracy using MRE (a) for cut-off frequency selection when p=0.5 and (b) for weighted-power selection when cut-off frequency is 4 Hz.*

Clean condition training	Set A	Set B	Set C	Avg.	Impr.
(1)MFCC(c0)	58.89	54.29	67.14	58.70	---------
(2)CMVN	77.52	78.86	78.53	78.26	---------
(3)CMVN+RASTA	77.70	79.00	78.41	78.36	0.45%
(4)CMVN+PCA(L=15)	80.31	81.15	82.02	80.99	12.56%
(5)CMVN+LDA(L=5)	81.54	82.65	82.85	82.25	18.35%
(6)CMVN+SHE	82.86	84.24	82.82	83.40	23.64%
(7)CMVN+MRE(best)	83.71	85.93	83.63	84.58	29.07%
(8)CMVN+SHE+MRE(best)	83.94	85.82	83.73	84.65	29.39%

Table 1: *Comparison of several representative methods for AURORA 2 clean-condition training. "Impr." is the error rate reduction as compared to CMVN.*

4.3. Analysis of Different Noise Types and Different SNR Values

In Fig 6 (a), we further compare the performance of the different approaches compared in Table 1 for different types of noise, but averaged over all SNR values. Every bar here in each set corresponds to a row in Table 1. We find that the proposed approaches (the last three bars (6) (7) (8)) performed better than the conventional approaches (bars (2) (4) (5)) for almost all types of noise.

In Fig 6 (b), we compare these methods for different SNR values but averaged over all noise types. Similar observations can be made. The proposed approaches (bars (6) (7) (8)) worked very well in all cases. In particular,

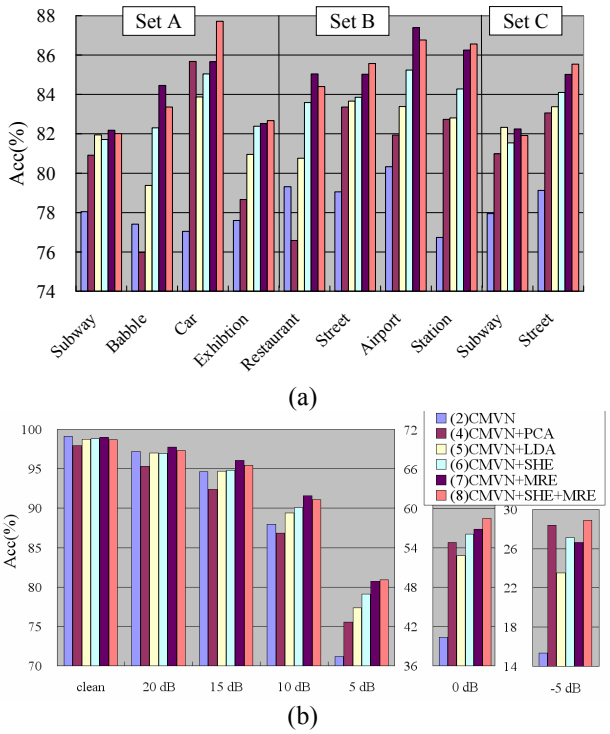


Figure 6: *Performance comparison for (a) different types of noise but averaged over all SNR values and (b) different SNR values but averaged over all types of noise.*

CMVN+MRE (bar (7)) offered almost the same performance as CMVN alone for clean speech, and was the best for the higher SNRs of 20, 15, and 10 dB. SHE+MRE (bar (8)) worked very well for lower SNRs (5, 0, and -5 dB), but turned out to be slightly worse than MRE alone for higher SNRs (20, 15, and 10 dB). This is probably because SHE involves equalization of the entire modulation spectrum distribution, so the speech characteristics may be over-fitted to a given distribution $CDF_{ref}[.]$ and thus some individual speech characteristics for each utterance may somehow be lost slightly, especially for higher SNR cases. MRE, on the other hand, only equalizes the magnitude ratio MR_d of the modulation spectrum and does not change many other statistics, and may therefore preserve more of the original speech characteristics. This may be the reason why MRE performed better than SHE for higher SNR cases.

4.4. Analysis of Time and Frequency Domain Behavior

When a modulation spectrum $Y_d(k)$ is transformed into another modulation spectrum $\hat{Y}_d(k)$, be it using SHE, MRE, or the combination thereof, there exists a corresponding "frequency response" $H_d(k)=\hat{Y}_d(k)/Y_d(k)$. The inverse Fourier transform of $H_d(k)$ gives the corresponding "impulse response" $h_d(n)$, although here $h_d(n)$ changes for each utterance. When $Y_d(k)$ is the modulation spectrum with CMVN alone, and $\hat{Y}_d(k)$ is that for modulation spectrum

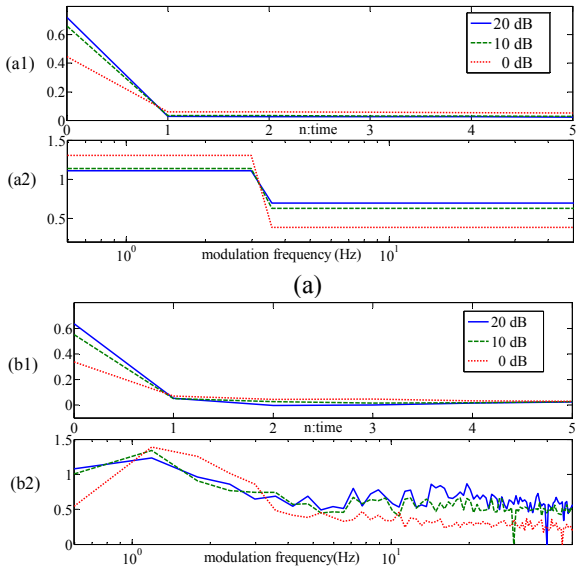


Figure 7: *Time and frequency domain behavior $h_d(n)$ and $H_d(k)$ for (a) MRE and (b)SHE+MRE.*

equalization (MRE or SHE+MRE) in addition, $h_d(n)$ and $H_d(k)$ then represent the time and frequency domain behavior of MRE or SHE+MRE. Figure 7 shows $h_d(n)$ and $H_d(k)$ for the c0 parameter for a typical example utterance in the AURORA 2 corpus respectively for MRE (Figure 7 (a)) and SHE+MRE (Figure 7 (b)) for the three SNR values 20, 10, and 0 dB. Figure 7 (a1) shows $h_d(n)$ for MRE. It can be found that for higher SNR, $h_d(n)$ is very close to a delta-Dirac function, that is, the temporal trajectories are essentially unchanged. With degradation of the SNR value, $h_d(n)$ becomes smoother; this represents the time domain behavior of MRE. The frequency domain behavior $H_d(k)$ of MRE is shown in Figure 7 (a2). We can observe that MRE acts as an ideal low-pass filter cut off at k_c for lower SNRs. For higher SNRs, MRE still acts like an ideal low-pass filter but the higher frequency parts are partially preserved while the lower frequency parts are slightly less enhanced. Intuitively, all of these trends are very reasonable. Shown in Figure 7 (b1) (b2) are the time and frequency domain behaviors of SHE+MRE; here similar observations can be made, except that with SHE the frequency domain behavior becomes slightly more complicated, as shown in Figure 7 (b2). Note that SHE and MRE actually adapt the filter coefficients (ie. $h_d(n)$ and $H_d(k)$) to different noise and SNR conditions for each utterance, so they performed very well in almost all conditions.

4.5. Integration of MRE with Other Feature Normalization Techniques

HEQ and higher-order cepstral moment normalization (HOCMN) [4] [15] have proved to be very useful for feature normalization in speech recognition tasks. Here we attempted integration of the MRE approach with these

techniques. In these experiments, the "feature normalization" block in Figure 4, previously represented by CMVN, was replaced by HEQ and HOCMN. We only consider MRE here because the additional improvements obtainable with SHE+MRE as shown in Table 1 were found to be limited, and indeed involved much higher computational costs. The results are shown in Table 2 for testing sets A, B, and C and the overall average. In row (1) are the results for CMVN, and serve as a reference, similar to row (2) in Table 1. Row (2) is the best result we obtained with HEQ on AURORA 2 using a progressive window with length l=98 frames, which is significantly better than CMVN. Row (3) is then HEQ followed by MRE, which offered a 11.07% relative error rate reduction with respect to HEQ (here we used cut-off frequency k_c=5 Hz and p=0.3 for MRE; these numbers were optimized empirically). Row (4) is the best result of HOCMN with AURORA 2 for integer order moments with the first, third, and 100-th order moments normalized [15], which is also significantly better than CMVN. Row (5) is HOCMN followed by MRE. A relative error rate reduction of 8.12% with respect to HOCMN was achieved (here we used cut-off frequency k_c=6 Hz and p=0.3 for MRE, also optimized empirically). Also listed in row (6) is the result for the advanced front-end (AFE) feature extraction algorithm recommended by ETSI [16], here used as a reference. We can see that the relatively simple HEQ+MRE or HOCMN+MRE are actually very close to, and in some cases higher than, the relatively complicated AFE. Note that when MRE is integrated with HEQ or HOCMN, a cut-off frequency of 5 Hz or 6 Hz turned out to be better than 4 Hz as used in section 4.1. This is reasonable because the noisy speech more closely resembles clean speech after performing HEQ or HOCMN, and thus higher cut-off frequencies should be used to preserve more information in the modulation spectrum. The results in Table 2 clearly show that the results of MRE when integrated with other feature normalization techniques are significantly better than those for CMVN alone, and similarly offer extra performance improvements.

Clean condition training	Set A	Set B	Set C	Avg.	Relative error rate reduction
(1)CMVN	77.52	78.86	78.53	78.26	---------
(2)HEQ	82.44	84.45	83.11	83.38	---------
(3)HEQ+MRE	84.31	86.47	84.56	85.22	(to HEQ) 11.07%
(4)HOCMN	83.78	86.12	83.87	84.73	---------
(5)HOCMN+MRE	85.10	87.15	85.34	85.97	(to HOCMN) 8.12%
(6)AFE	86.49	85.58	84.90	85.81	---------

Table 2: *Recognition results of MRE integrated with HEQ and HOCMN under AURORA 2 clean-condition training.*

5. CONCLUSION

We proposed a new method for generating robust features using modulation spectrum equalization. The techniques of spectral histogram equalization (SHE) and magnitude ratio equalization (MRE) can offer significant improvements for different types of noise and different SNR values. We also showed that the proposed approach can be integrated with CMVN or other more advanced feature normalization techniques. These results indicate the effectiveness of equalization performed on the modulation spectrum in reducing the mismatch produced by additive and convolutional noise.

6. REFERENCES

[1] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on ASSP, 1981.

[2] O.Viikki, K.Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition", Speech Communication, August 1998.

[3] A.de la Torre, A.M.Peinado, J.C.Segura, J.L.Perez-Cordoba, M.C.Benitez, A.J.Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition", IEEE Trans. on SAP, May 2005.

[4] C-W.Hsu, L-S.Lee, "Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition", ICASSP 2004.

[5] H.Hermansky and N.Morgan, " RASTA processing of speech", IEEE Trans. on SAP, 1994.

[6] H.Hermansky and P.Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR", Interspeech 2005.

[7] S. van Vuuren and H. Hermansky, "Data-driven Design of RASTA-like Filters", Eurospeech 1997.

[8] J-W.Hung, H-M.Wang, and L-S.Lee, "Comparative Analysis for Data-driven Temporal Filters Obtained via Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) in Speech Recognition", Eurospeech 2001.

[9] N-C.Wang, J-W.Hung, L-S.Lee, "Data-driven Temporal Filters Based on Multi-eigenvectors for Robust Features in Speech Recognition", ICASSP 2003.

[10] J-W.Hung ,L-S.Lee, "Optimization of Temporal Filters for Constructing Robust Features in Speech Recognition", IEEE Trans. on SAP, May 2006.

[11] N.Kanedera, T.Arai, H.Hermansky, and M.Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Communication, 1999.

[12] V.Tyagi, I.McCowan, H.Misra, H.Bourlard, "Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR", IEEE ASRU 2003.

[13] X.Xiao, E.S.Chng, H.Li, " Normalizing the Speech Modulation Spectrum for Robust Speech Recognition", ICASSP 2007.

[14] H.G.Hirsch, D.Pearce," The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, September 2000.

[15] C-W.Hsu, L-S.Lee, "Extension and Further Analysis of Higher Order Cepstral Moment Normalization (HOCMN) for Robust Features in Speech Recognition", Interspeech 2006.

[16] ETSI ES 202 212 v1.1.1, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm; back-end speech reconstruction algorithm", Nov. 2003.