Evaluation of Modulation Spectrum Equalization Techniques for Large Vocabulary 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

Previous approaches for modulation spectrum equalization were evaluated only for the Aurora 2 small vocabulary task. We further apply these approaches on the Aurora 4 large vocabulary task. 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. Experimental test results indicate significant performance improvements using these approaches when cascaded with cepstral mean and variance normalization (CMVN). Cascading MRE with more advanced feature normalization approaches such as histogram equalization (HEQ) and higher-order cepstral moment normalization (HOCMN) yielded additional performance improvements.

Index Terms: temporal filtering, modulation spectrum, feature normalization, robust feature extraction

1. Introduction

The inevitable mismatch between training and testing conditions very often seriously degrades the performance of statistical speech recognition approaches. Many feature-based approaches have been proposed to handle this problem and they can roughly be represented as the framework in Figure 1.

The first block is the feature normalization technique, which tries to normalize the statistical parameters of the speech features in order to reduce the effect of 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],[5] 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, normalizes 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 in the cepstrum, although in different ways.

Typical examples of post processing in the second block include filtering the time trajectories or the modulation spectrum of the speech features, among many others. These include RASTA filtering [6],[7], temporal smoothing filter (TES)[8], square-root Wiener filtering on the modulation spectrum [9],[10], with filters designed by data-driven methods based on different criteria such as linear discriminant analysis (LDA) [11], principle component analysis (PCA) [12], and minimum classification error (MCE) [13], and so on.

In our recent work, we proposed two modulation spectrum equalization techniques [14] to reduce the mismatch between the training and testing environments. The first is to equalize the cumulative density functions (CDFs) of the modulation spectra of clean and noisy speech, and the second is to equalize the magnitude ratio of lower to higher components in the modulation spectrum. Preliminary experiments performed on the small vocabulary English-digit/s AURORA 2 test environments offered very encouraging results.

In this paper, we try to evaluate these two approaches of modulation spectrum equalization on the Aurora 4 large vocabulary task. We also compare the performance of modulation spectrum equalization techniques with other well-known temporal filtering approaches. The rest of the paper is organized as follows. In section 2, we briefly introduce the proposed modulation spectrum equalization techniques. In sections 3 and 4 the experimental setup and results are reported. Concluding remarks are finally presented in section 5.

2. Modulation Spectrum Equalization Techniques

Given a sequence of feature vectors \( \{x(n), n=1,2, ..., N\} \) for an utterance, each including \( D \) feature parameters,

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

where \( n \) is the time index, and \( d=1,...,D \) is the parameter index. 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 \( y_d(n) = x(n,d) \). The modulation spectrum \( Y_d(k) \) of the \( d \)-th time trajectory can be obtained by applying discrete Fourier transform [13]:

\[
Y_d(k) = \sum_{n=0}^{N-1} y_d(n) \exp(-j \pi nk / N),
\]

\( k = 0,1,2, ..., N-1; \quad d = 1,2, ..., D \),

where \( k \) is the frequency index of the discrete Fourier transform. The two techniques proposed here can then be performed with \( Y_d(k) \). 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. To reduce the computation complexity, we use the FFT algorithm instead of the discrete
Fourier transform; that is, if N is not 2\(^n\) (m is a positive integer), we pad zeros to \(y_d(n)\) such that N will be 2\(^n\).

### 2.1. Spectral Histogram Equalization (SHE)

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 4 to be used as the reference CDF, CDF\(_{\text{ref}}\). For any test utterance, let \(Y_{d,test}(k)\) represent the modulation spectrum of the \(d\)-th time trajectory of a test utterance. The CDF for this modulation spectrum magnitude, \(|Y_{d,test}(k)|\), can be similarly obtained as CDF\(_{\text{test}}\). The concept of spectral histogram equalization is plotted in Figure 2, where the blue curve is CDF\(_{\text{test}}\) and the red curve is the reference CDF, CDF\(_{\text{ref}}\). Hence, we can map \(|Y_{d,test}(k)|\) to the equalized magnitude \(|\hat{Y}_{d,test}(k)|\) by:

\[
|\hat{Y}_{d,test}(k)| = \text{CDF}^{-1}_\text{ref}(\text{CDF}_\text{test}[|Y_{d,test}(k)|]) ,
\]

where CDF\(_{-1}\) 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.

### 2.2. Magnitude Ratio Equalization (MRE)

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

\[
MR_d = \frac{\sum_{k=1}^{N/2} |Y_{d}(k)|}{\sum_{k=N/2+1}^{N} |Y_{d}(k)|} ,
\]

where \(k_c\) is the cut-off frequency whose value can be determined empirically. \(N\) is the order of the FFT, and \([N/2]\) is a function that returns the largest integer less than or equal to \(N/2\). 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. Our analysis showed that the value of MR\(_d\) is highly correlated with SNR values [14]. 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 4 as the reference value MR\(_{\text{ref}}\). Likewise, we then calculate the value of MR\(_d\) for each test utterance as MR\(_{\text{test}}\). We then equalize the magnitude of the modulation spectrum for the test utterance \(|Y_{d,test}(k)|\) as:

\[
|\hat{Y}_{d,test}(k)| = \left\{ \begin{array}{ll} 
\frac{\text{MR}_d}{\text{MR}_{\text{ref}}} \cdot |Y_{d,test}(k)| & , \ k \leq k_c \\
1 - \frac{\text{MR}_d}{\text{MR}_{\text{ref}}} (1-p) & , \ k > k_c 
\end{array} \right. 
\]

where 0<p<1 is the weighted-power for the scaling factor. For example, if p=0.3, we make the value of MR\(_{\text{test}}\) identical to MR\(_{\text{ref}}\), but the lower frequency components (or speech) are less enhanced while the higher frequency components (or noise) are more suppressed. The best value of \(p\) here can again be determined empirically.

### 2.3. The Framework of the Proposed Approaches

The framework of the proposed approach is to cascade the popular feature normalization techniques with the modulation spectrum equalization techniques proposed here as shown in Figure 1. 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. We then perform SHE and MRE on the modulation spectrum, and transform the features back to the cepstral domain by inverse FFT. Note that each test utterance has its own modulation spectrum histogram CDF\(_{\text{test}}\) and MR\(_{\text{test}}\) values, and thus is transformed individually. Hence, SHE and MRE can be considered as adaptive filters which can adapt the filter coefficients to different speech and noise conditions. This is different from many conventional temporal filtering approaches, in which the same transformation (or same set of filter coefficients) is used for different utterances and different noise conditions.

### 3. Experimental Setup

The above approaches were evaluated under the AURORA 4 [15] test environment, which is derived from the Wall Street Journal (WSJ0) 5k-words dictation task. In the Aurora 4 corpus, speech data were sampled in both 8k Hz and 16k Hz, and we only used the 8k Hz speech data in this work. We built the baseline system with cross-word triphone HMM via Hidden Markov Model Toolkit (HTK), in which the phone sets and lexicon were based on those provided by Carnegie Mellon University [16]. The baseline acoustic models were trained from 7138 clean training utterances (about 12 hours), and there were 3 emitting states in each triphone HMM with 8 Gaussian mixtures per state. Totally there were around 2960 tied states and 23000 Gaussian mixtures in our experiments. A bigram language model for a 5k-word closed vocabulary provided by MIT Lincoln Laboratory was used for test, while the decoding was done by HVite in HTK, and the language model weight, the word insertion penalty, and the pruning factor were respectively set to 15,-4, and 250.

There are 14 test sets in the Aurora 4 task, and each test set has 166 test utterances (small test set defined in [15]). The test sets 01-07 were recorded by a close-talk microphone which was the same as the microphone used for the clean training data, and the test sets 08-14 were recorded with a microphone selected from a set of different microphones. Each test utterance was corrupted by one type of additive noise with SNR values between 5 dB to 15 dB (average SNR value is 10dB). The test set 01 is the clean speech, and set 02-07 are corrupted by car, babble, restaurant, street, airport, and train noise. The additive noises in test set 08-14 are the same as set 01-07, except they were recorded by different microphones.

The speech features were extracted by the AURORA W1007 front-end, which converted each signal frame into 13 cepstral coefficients (MFCCs, c0-c12), on which the feature
normalization techniques and all the modulation spectrum equalization techniques mentioned 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. Performance of Modulation Spectrum Equalization Approaches on the AURORA 4 Task

The experimental results for clean condition training were shown in Table 1. We first tested the several feature normalization techniques and used them as the baseline for comparison for the modulation spectrum equalization techniques analyzed here. This included MFCC baseline (row (1)), CMVN (row (2)), HEQ (row (7)), and the previously proposed Higher Order Cepstral Moment Normalization (HOCMN) (row (9)) respectively. In all cases, we set the parameters used in these techniques to be the same as those we used in Aurora 2. For HOCMN the first, third, and 100-th order moments were normalized. From the average results in the second last column, we see HOCMN (row (9)) offered absolute averaged WER reduction of 4.02% and 1.50% over CMVN (row (2)) and HEQ (row (7)) respectively. Hence, the results of HOCMN for large vocabulary task were consistent with our previous paper evaluated in the Aurora 2 task [4][5].

We then evaluated the performance of the various post processing techniques when applied in addition to the normalization techniques. The last column of Table 1 shows the relative word error rate reduction with respect to the corresponding feature normalization techniques alone. Rows (3) and (4) are CMVN followed by temporal filtering based on LDA criterion [10][11] (with optimal filter length L=5 trained on 41 monophone classes) and PCA criterion [11] (with optimal filter length L=5), and they actually couldn’t offer improvements over CMVN in the large vocabulary task here, although they performed very well in Aurora 2 task (see Table 1 in [14]). However, CMVN+SHE (row (5)) and CMVN+MRE (row (6), kc=6 Hz, p=0.2) presented in this paper provided significant improvements over CMVN alone, respectively offering 2.99% and 5.12% relative word error rate reduction. Moreover, CMVN+MRE (row (6)) was better than CMVN+SHE (row (5)) in almost all cases listed in the table and used less computational cost, so in many cases we may choose MRE for simplicity.

We then integrated MRE with HEQ and HOCMN. It can be found that HEQ+MRE (row (8), kc=5 Hz, p=0.3) and HOCMN+MRE (row (10), kc=7 Hz, p=0.2) respectively reduced the averaged WER by 7.46% and 3.15% with respect to HEQ and HOCMN alone. Also listed in row (11) is the result for the advanced front-end (AFE) feature extraction algorithm recommended by ETSI [17], here used as a reference. We can see that the relatively simple HEQ+MRE or HOCMN+MRE were actually very close to, and in some cases the table better than the AFE.

For ease of comparison of these approaches, we also averaged the word accuracies of sets 02-07 and sets 09-14 as the performance measure respectively for “additive noise” and “channel mismatch + additive noise”, and then plotted them in Figure 3. Every bar here in each set corresponds to a row in Table 1. Similar observation can be found as in Table 1. The proposed approaches (bars (5) (6) (8) (10)) offered reasonable improvements over the conventional feature normalization techniques.

4.2. Discussion

From the experimental results, some differences between the large vocabulary task AURORA 4 and small vocabulary Aurora 2 task can be observed. In general, the temporal filtering and modulation spectrum equalization all tried to smooth the feature parameters in the time trajectories or to remove the noise in the higher modulation frequencies. However, speech information contained in the higher modulation frequencies may be distorted after we performed such filtering. There are only 11 word models in Aurora 2 task, so the filtered features remain discriminative enough for recognition purposes although some of the speech information in the higher

![Figure 3: Performance comparison for the several approaches considered. The vertical axis is the word accuracies averaged over sets 02-07 (for “additive noise”) and 09-14 (for “channel mismatch + additive noise”.

Table 1. Recognition results of the modulation spectrum equalization approaches under AURORA 4 test environments. “Avg.” is the averaged word accuracy for all test sets (set 01-14). “Impr.” is the relative word error rate reduction as compared to the corresponding feature normalization techniques alone, i.e. “Impr.” of rows (3)-(6) are compared to CMVN alone, while rows (8) and (10) are compared to HEQ and HOCMN alone.
modulation frequencies may be distorted. However, when the number of models to be distinguished becomes large in LVCSR, the features should be less filtered so that they can preserve more discriminative speech information for recognition. This may be the reason why the conventional temporal filtering approaches (PCA or LDA) performed very well in small vocabulary task, but worse in LVCSR. Also, according to our observations in Aurora 2, these techniques performed very well only in lower SNR values (-5 dB to 5 dB), but worse in higher SNR values. Because they used the same transformation for all test conditions, the features may be over smoothed for clean speech data so as to lose their speech information. However, SHE and MRE are different. They can adapt the filter coefficients for different noisy conditions, so they can retain more speech information for higher SNR cases by mild smoothing of the features. This may be the reason why our proposed approaches performed better than the conventional temporal filtering approaches, especially in the large vocabulary task. Note that the best cut-off frequencies $k_c$ of MRE used in Aurora 4 are 5-7 Hz, which are a little higher than what we used in Aurora 2 (4-6 Hz). This is also reasonable since we need to preserve more discriminative speech information in the modulation spectrum for LVCSR.

4.3. Further Analysis of the Modulation Spectrum Equalization Approaches via Distance Measure

For further analysis of the proposed approaches, we define the averaged distance measure $l$:

$$l = E\left[\frac{1}{N} \sum_{i=1}^{N} \|x_i - y_i\|_2^2\right]$$

(6)

where $x$ is the 13-dimensional vector of MFCC parameters for clean speech and $y$ is the corresponding noisy speech version processed by some feature normalization and/or post processing approaches. $\|\|$ is the Euclidean distance, and the average $E[.]$ is performed over all utterances in the test set. Therefore the distance measure $l$ reflects in average how the equalized feature vectors are matched to their clean speech versions individually. The results of the distance measure $l$ evaluated in Aurora 4 are listed in Table 2, in which we averaged $l$ over the sets 02-07 and 09-14 respectively for cases of “additive noise” and “channel mismatch + additive noise”. Row (2) and row (3) show that CMVN+SHE and CMVN+MRE clearly reduced the distance between the noisy speech and clean speech as compared to CMVN alone. Also listed in row (5) and row (7) are the distance measure of HEQ+MRE and HOCMN+MRE, also reduced as compared to HEQ alone or HOCMN alone. Also, these distance measures actually have close correlation with the accuracies listed in Table 1 and shown in Figure 3, which indicates these distances are meaningful.

### Table 2. Comparison of distance measure $l$ (defined in equation (6)) of the approaches analyzed here. We averaged over all test utterances in the sets 02-07 and 09-14 for “additive noise” and “channel mismatch + additive noise” respectively.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>additive noise</th>
<th>channel mismatch + additive noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CMVN</td>
<td>0.9314</td>
<td>0.9886</td>
</tr>
<tr>
<td>2CMVN+SHE</td>
<td>0.9208</td>
<td>0.9614</td>
</tr>
<tr>
<td>3CMVN+MRE</td>
<td>0.9156</td>
<td>0.9477</td>
</tr>
<tr>
<td>4HEQ</td>
<td>0.9085</td>
<td>0.9435</td>
</tr>
<tr>
<td>5HEQ+MRE</td>
<td>0.8821</td>
<td>0.9081</td>
</tr>
<tr>
<td>6HOCMN</td>
<td>0.9051</td>
<td>0.9330</td>
</tr>
<tr>
<td>7HOCMN+MRE</td>
<td>0.8880</td>
<td>0.9054</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, we evaluated the spectral histogram equalization (SHE) and magnitude ratio equalization (MRE) techniques on the large vocabulary Aurora 4 task, and compared them with several conventional temporal filtering approaches. Our results showed that several conventional temporal filtering approaches can’t improve the performance in LVCSR although they performed very well in small vocabulary task. However, the proposed approach of SHE and MRE can be integrated with CMVN or other more advanced feature normalization techniques to improve the performance in LVCSR. These results indicate the effectiveness of equalization performed on the modulation spectrum in reducing the mismatch produced by additive and convoluted noise.

6. Acknowledgements

This work is supported by the National Taiwan University Advance Speech Technology Scholarship.

7. References