Evaluation of Modulation Spectrum Equalization Techniques for Large Vocabulary Robust Speech Recognition

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

Wen-Yi Chu

Outline

- Introduction
- Modulation Spectrum Equalization Techniques
- Experimental Setup
- Experimental Results
- Conclusions

Introduction

- 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 our recent work, we proposed two modulation spectrum equalization techniques 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.

- 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.

Modulation Spectrum Equalization Techniques

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

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

- The modulation spectrum $Y_d(k)$ of the d-th time trajectory can be obtained by applying discrete Fourier transform :

$$Y_d(k) = \sum_{n=0}^{N-1} y_d(n) \cdot \exp(-j\pi nk / N), \qquad (2)$$

$$k = 0,1,2,\ldots, N-1; \quad d = 1,2,\ldots, D$$

where k is the frequency index of the discrete Fourier transform.

- 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.

Modulation Spectrum Equalization Techniques

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, $CDF_{ref}[\cdot]$.
- We can map $\left|Y_{d,test}(k)\right|$ to the equalized magnitude $\left|\hat{Y}_{d,test}(k)\right|$ by:

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

where $CDF^{-1}{}_{ref}[\cdot]$ is the inverse of the cumulative distribution function.

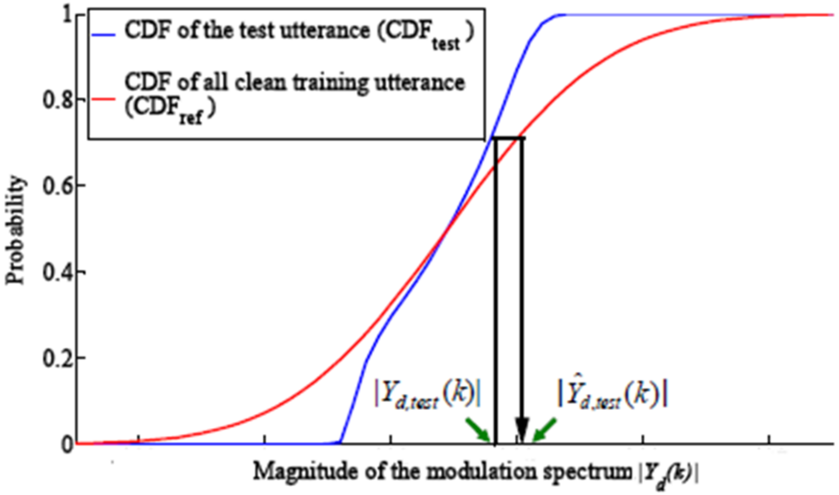


Figure 2: *The concept of the spectral histogram equalization (SHE).*

Modulation Spectrum Equalization Techniques

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:

$$MR_d = \frac{\sum_{k=0}^{k_c} \left|Y_d(k)\right|}{\sum_{k=k_c+1}^{[\frac{N}{2}]+1} \left|Y_d(k)\right|} \qquad (4)$$

where k_c is the cut-off frequency whose value can be determined empirically, N is the order of the FFT.

- We then equalize the magnitude of the modulation spectrum for the test utterance $\left|Y_{d,test}(k)\right|$ *as*

$$\left|\hat{Y}_{d,test}(k)\right| = \begin{cases} \left(\frac{MR_{d,ref}}{MR_{d,test}}\right)^{p} \cdot \left|Y_{d,test}(k)\right| & ,k \le k_c \\ \frac{1}{\left(\frac{MR_{d,ref}}{MR_{d,test}}\right)^{1-p}} \cdot \left|Y_{d,test}(k)\right| & ,k > k_c \end{cases} \qquad (5)$$

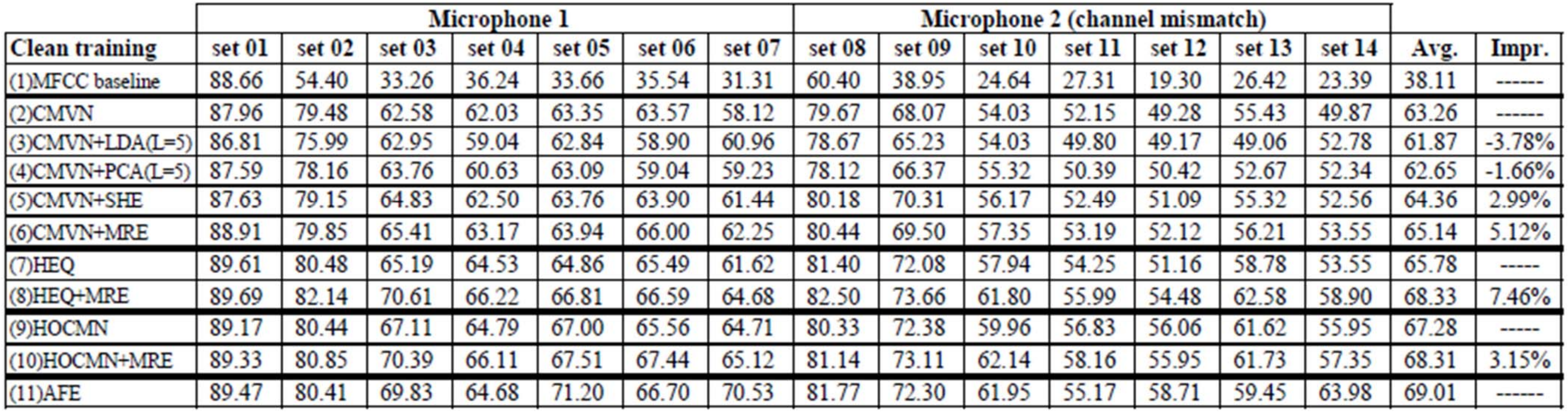
Experimental Setup

- The above approaches were evaluated under the AURORA 4 test environment, which is derived from the Wall Street Journal (WSJ0) 5k-words dictation task.
- 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.

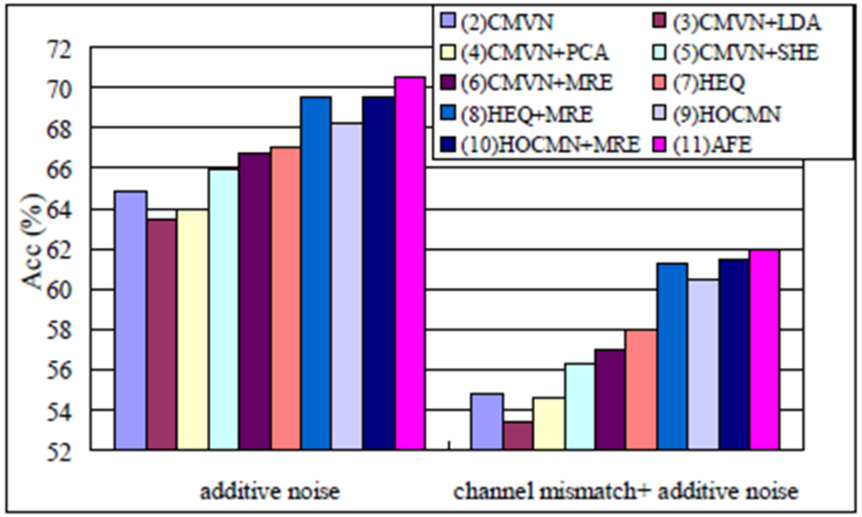
	set 01	se t 02	set 03	set 04	set 05	set 06	set 07
Microphone	Microphone 1(as training data)						
additive noise	clean	car	babble	Restau rant	street	airport	train

	set 08	se t 09	set 10	set 11	set 12	set 13	set 14
Microphone	Microphone 2(channel mismatch)						
additive noise	clean	car	babble	Restau rant	street	airport	train

Experimental Results



	Microphone 1							Microphone 2 (channel mismatch)								
Clean training	set 01	set 02	set 03	set 04	set 05	set 06	set 07	set 08	set 09	set 10	set 11	set 12	set 13	set 14	Avg.	Impr.
(1)MFCC baseline	88.66	54.40	33.26	36.24	33.66	35.54	31.31	60.40	38.95	24.64	27.31	19.30	26.42	23.39	38.11	------
(2)CMVN	87.96	79.48	62.58	62.03	63.35	63.57	58.12	79.67	68.07	54.03	52.15	49.28	55.43	49.87	63.26	------
(3)CMVN+LDA(L=5)	86.81	75.99	62.95	59.04	62.84	58.90	60.96	78.67	65.23	54.03	49.80	49.17	49.06	52.78	61.87	-3.78%
(4)CMVN+PCA(L=5)	87.59	78.16	63.76	60.63	63.09	59.04	59.23	78.12	66.37	55.32	50.39	50.42	52.67	52.34	62.65	-1.66%
(5)CMVN+SHE	87.63	79.15	64.83	62.50	63.76	63.90	61.44	80.18	70.31	56.17	52.49	51.09	55.32	52.56	64.36	2.99%
(6)CMVN+MRE	88.91	79.85	65.41	63.17	63.94	66.00	62.25	80.44	69.50	57.35	53.19	52.12	56.21	53.55	65.14	5.12%
(7)HEQ	89.61	80.48	65.19	64.53	64.86	65.49	61.62	81.40	72.08	57.94	54.25	51.16	58.78	53.55	65.78	-----
(8)HEQ+MRE	89.69	82.14	70.61	66.22	66.81	66.59	64.68	82.50	73.66	61.80	55.99	54.48	62.58	58.90	68.33	7.46%
(9)HOCMN	89.17	80.44	67.11	64.79	67.00	65.56	64.71	80.33	72.38	59.96	56.83	56.06	61.62	55.95	67.28	-----
(10)HOCMN+MRE	89.33	80.85	70.39	66.11	67.51	67.44	65.12	81.14	73.11	62.14	58.16	55.95	61.73	57.35	68.31	3.15%
(11)AFE	89.47	80.41	69.83	64.68	71.20	66.70	70.53	81.77	72.30	61.95	55.17	58.71	59.45	63.98	69.01	------



Experimental Results

Discussion

- 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 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.

- SHE and MRE 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.

Experimental Results

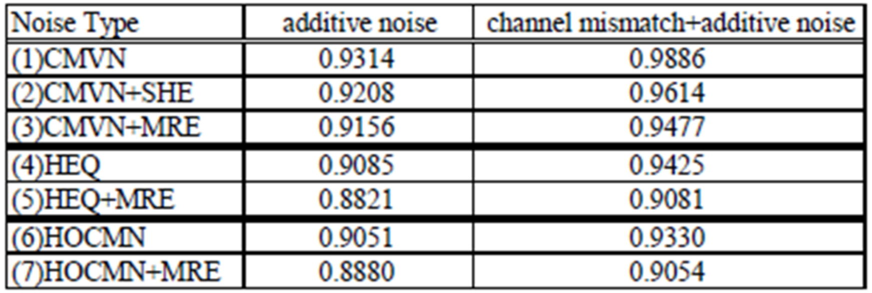
Further Analysis of the Modulation Spectrum Equalization Approaches via Distance Measure

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

$$l = E\left[\frac{\|\bar{y}-\bar{x}\|}{\|\bar{x}\|}\right] \qquad (6)$$

where $\bar{x}$ is the 13-dimensional vector of MFCC parameters for clean speech and $\bar{y}$ is the corresponding noisy speech version but processed by some feature normalization and/or post processing approaches, $\|\cdot\|$ is the Euclidean distance, and the average E[.] is performed over all utterances in the test set.

- These distance measures actually have close correlation with the accuracies listed in Table 1 and shown in Figure 3, which indicates these distance measures are meaningful.



Noise Type	additive noise	channel mismatch+additive noise
(1)CMVN	0.9314	0.9886
(2)CMVN+SHE	0.9208	0.9614
(3)CMVN+MRE	0.9156	0.9477
(4)HEQ	0.9085	0.9425
(5)HEQ+MRE	0.8821	0.9081
(6)HOCMN	0.9051	0.9330
(7)HOCMN+MRE	0.8880	0.9054

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.

- 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.

- The results indicate the effectiveness of equalization performed on the modulation spectrum in reducing the mismatch produced by additive and convolutional noise.