MODULATION SPECTRUM EQUALIZATION FOR ROBUST SPEECH RECOGNITION

Source: Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on

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

Professor : 陳嘉平

Reporter : 楊治鏞

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.
- 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.

Introduction

- 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.

Modulation spectrum (1/2)

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

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

- where *n* is the time index, and $d = 1, \ldots, D$ is the parameter index.

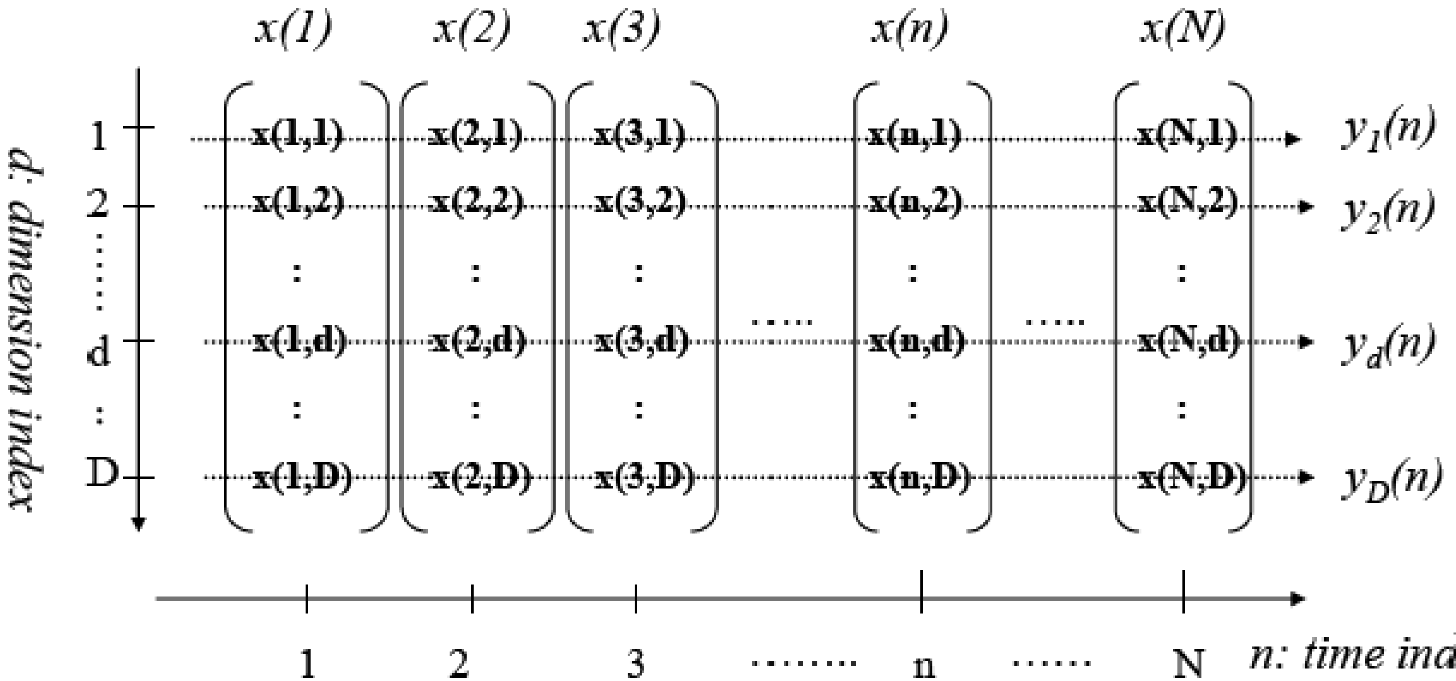


Figure 1: *The representation of the time trajectories of feature parameter sequences*

Modulation spectrum (2/2)

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

$$Y_d(k) = \sum_{n=0}^{N-1} y_d(n) \cdot \exp(-j2\pi nk / N)$$

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

Spectral Histogram Equalization

- We first calculate the **cumulative distribution function** (CDF) of the magnitudes of the modulation spectra, $\left|Y_d(k)\right|$, 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, $\left|Y_{d,test}(k)\right|$, can be similarly obtained as $\mathrm{CDF}_{\mathrm{test}}[\cdot]$.

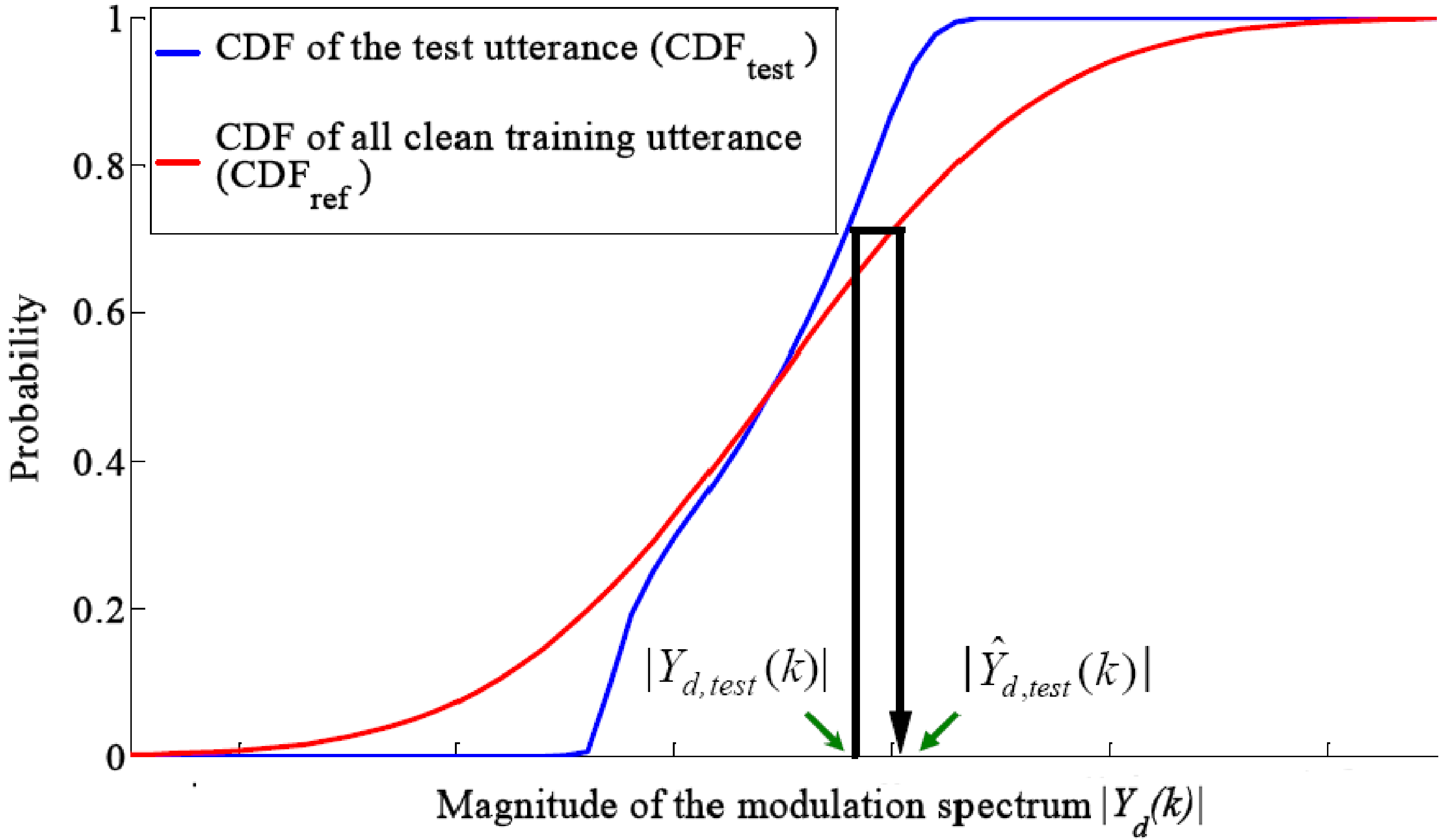


Figure 2: *The concept of the spectral histogram equalization (SHE).*

Spectral Histogram Equalization

- Hence the equalized magnitude of modulation spectrum $\left|\hat{Y}_{d,test}(k)\right|$ is

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

Magnitude Ratio Magnitude Ratio Equalization

- We first define a magnitude ratio (MR) for lower to higher frequency components for each parameter index d as follows:

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

- where k_c is the cut-off frequency used here, N is the order of the discrete Fourier transform.

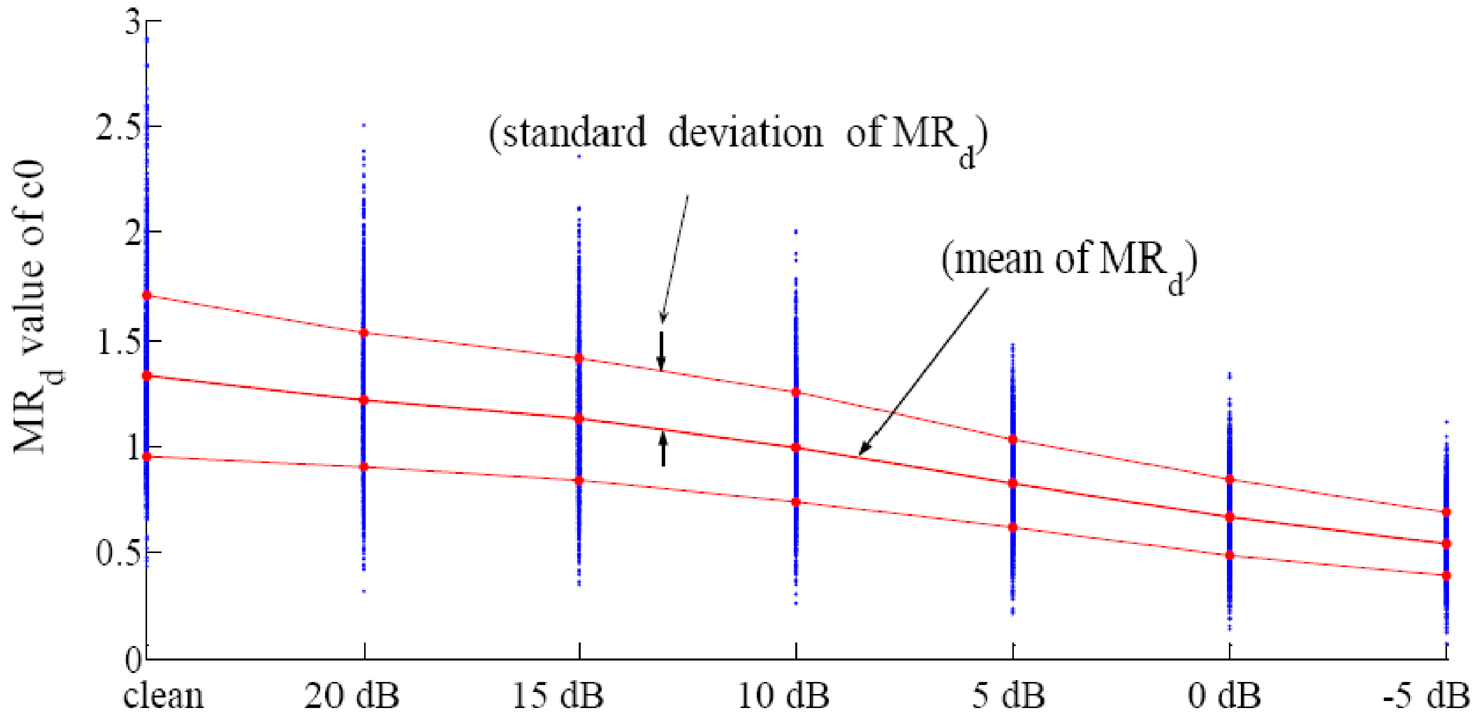


Figure 3: *The distribution of the magnitude ratio (*MR$_d$*) values of c0 for all testing utterances in AURORA 2 for all sets at all SNRs. Each point represents the* MR$_d$ *value of c0 for an utterance.*

Magnitude Ratio

- 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.

Magnitude Ratio Equalization

- We first calculate the average of MR_d for all utterances in the clean training data of AURORA 2 as the reference value $MR_{d,ref}$.

- We then calculate the value of MR_d for each test utterance as $MR_{d,test}$.

Magnitude Ratio Equalization

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

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

where $0 < p < 1$ is the weighted-power for the scaling factor.

EXPERIMENTAL SETUP

- AURORA 2
- The speech features were extracted by the AURORA WI007 front-end.
- Figure 4

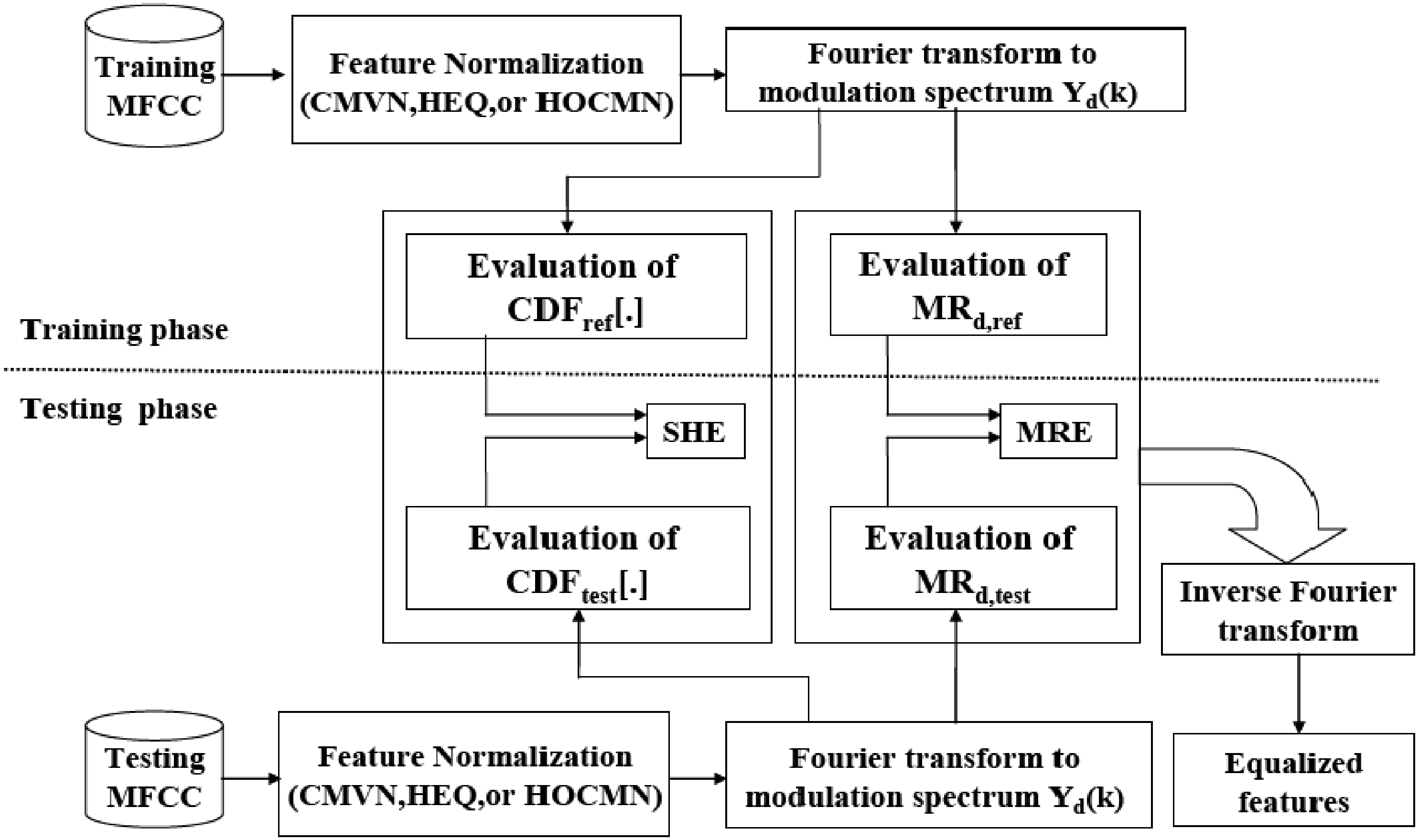
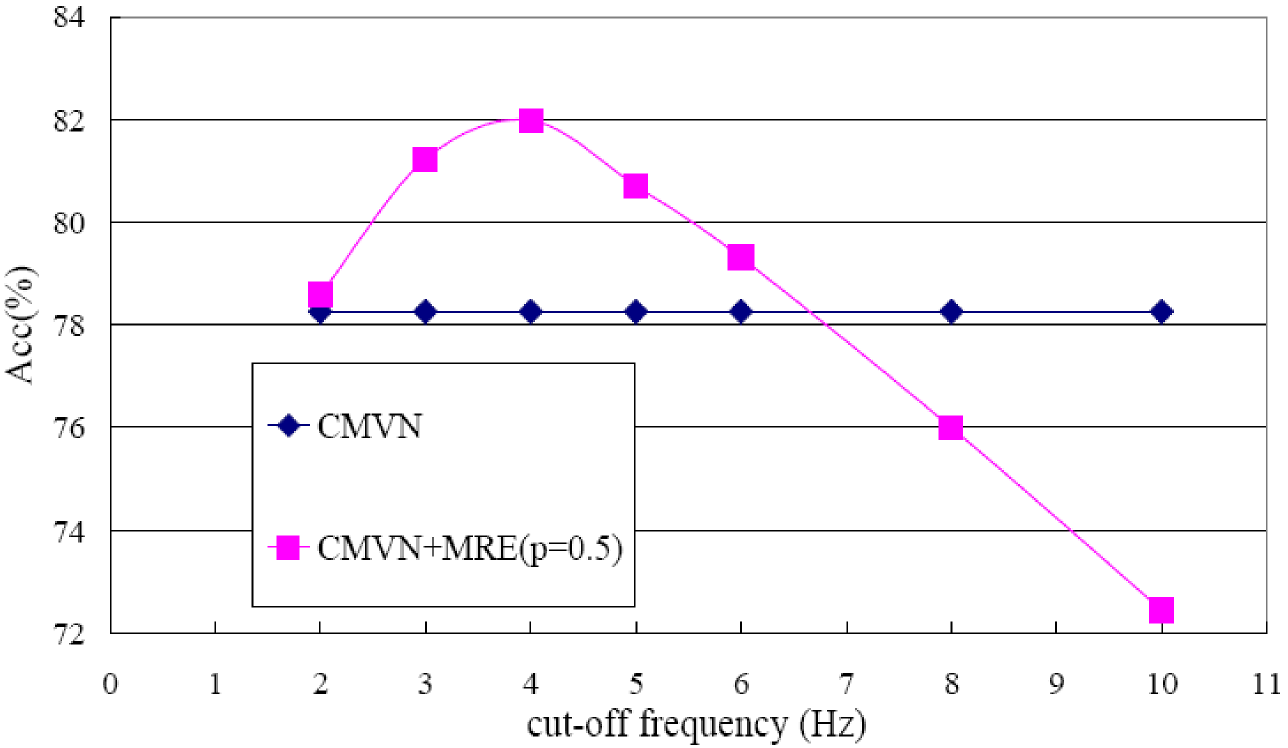
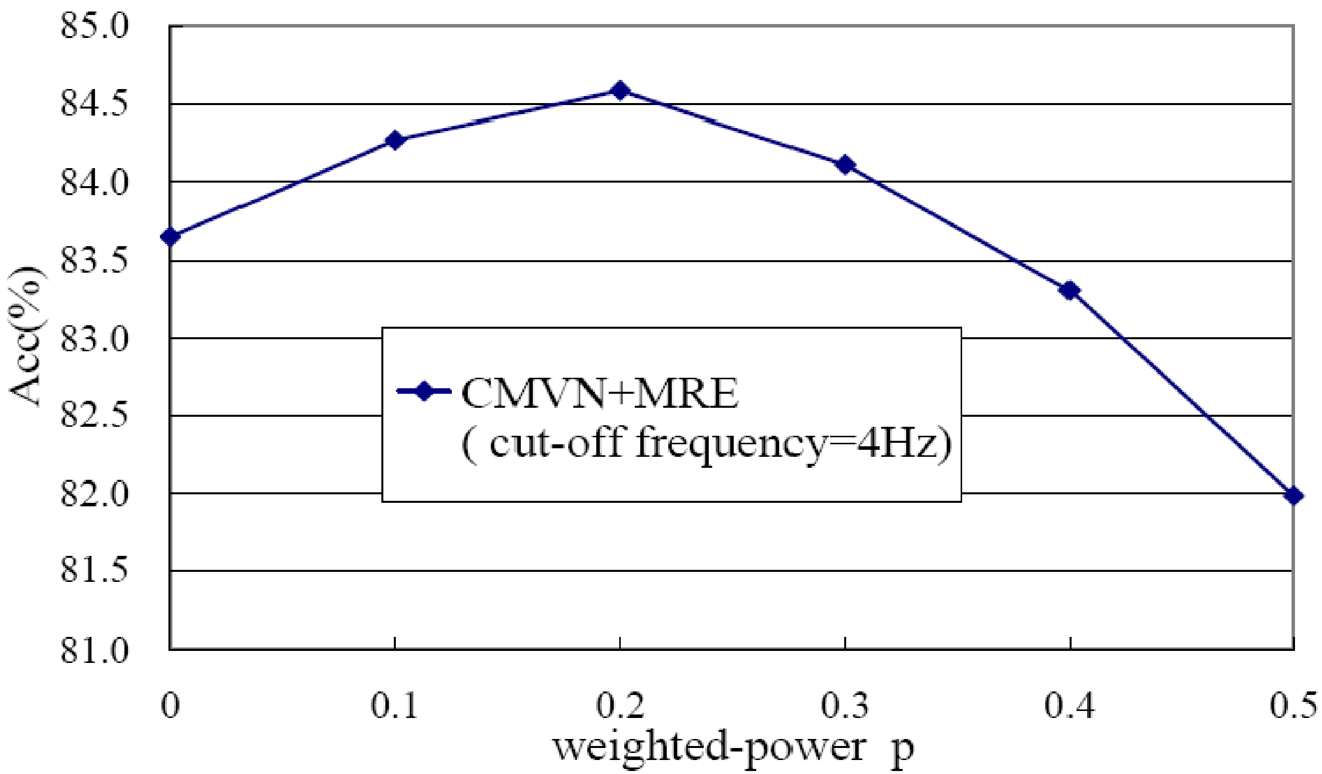


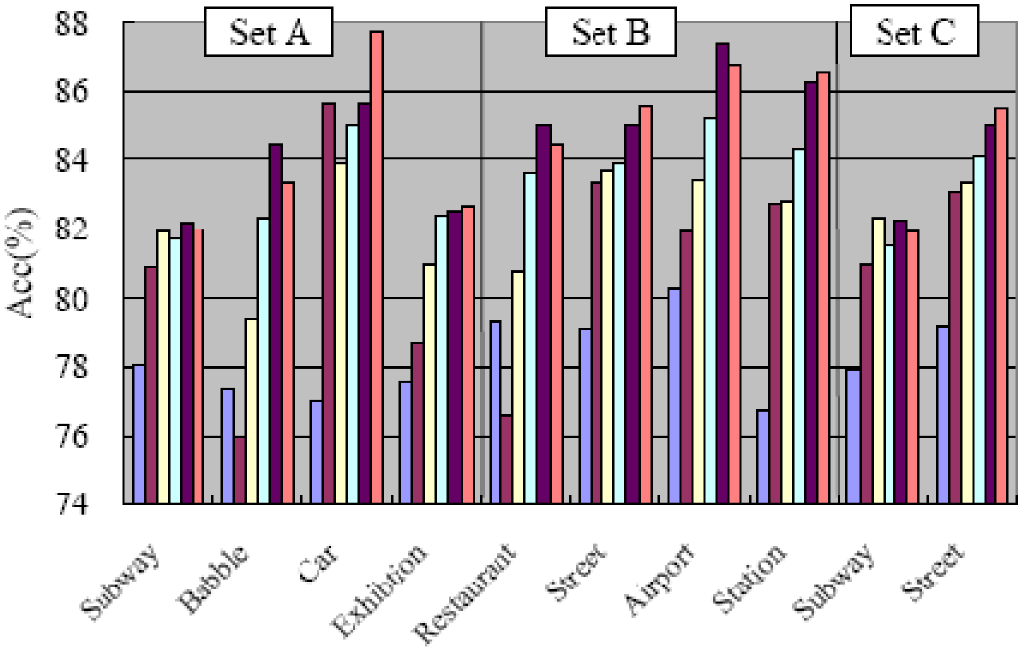
Figure 4: *The overall framework of modulation spectrum equalization techniques.*


84
82
80
78
76
74
72
Acc(%)
0
1
2
3
4
5
6
7
8
9
10
11
cut-off frequency (Hz)
CMVN
CMVN+MRE(p=0.5)

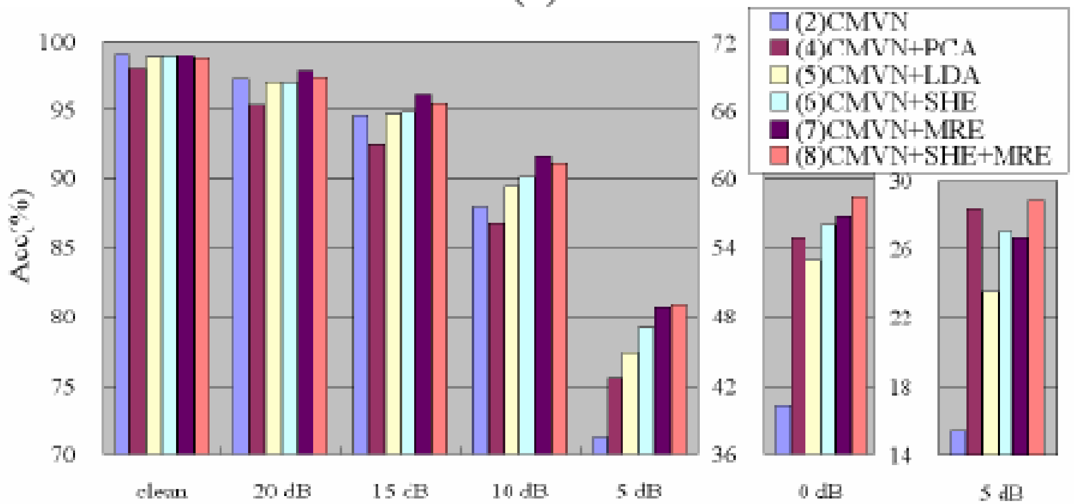

85.0
84.5
84.0
83.5
83.0
82.5
82.0
81.5
81.0
Acc(%)
CMVN+MRE
(cut-off frequency=4Hz)
0
0.1
0.2
0.3
0.4
0.5
weighted-power p

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.*


Set A
Set B
Set C
Acc(%)
88
86
84
82
80
78
76
74
Subway
Babble
Car
Exhibition
Restaurant
Street
Airport
Station
Subway
Street

(a)


(2)CMVN
(4)CMVN+PCA
(5)CMVN+LDA
(6)CMVN+SHE
(7)CMVN+MRE
(8)CMVN+SHE+MRE
Acc(%)
100
95
90
85
80
75
70
72
66
60
54
48
42
36
30
26
22
18
14
clean
20 dB
15 dB
10 dB
5 dB
0 dB
-5 dB

(b)

Integration of MRE with Other Feature Normalization Techniques

- 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.

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.*