Cepstrum-Domain Model Combination Based on Decomposition of Speech and Noise Using MMSE-LSA for ASR in Noisy Environments

Author : Hong Kook Kim and Richard C. Rose

Professor: 陳嘉平

Reporter: 吳國豪

Outline

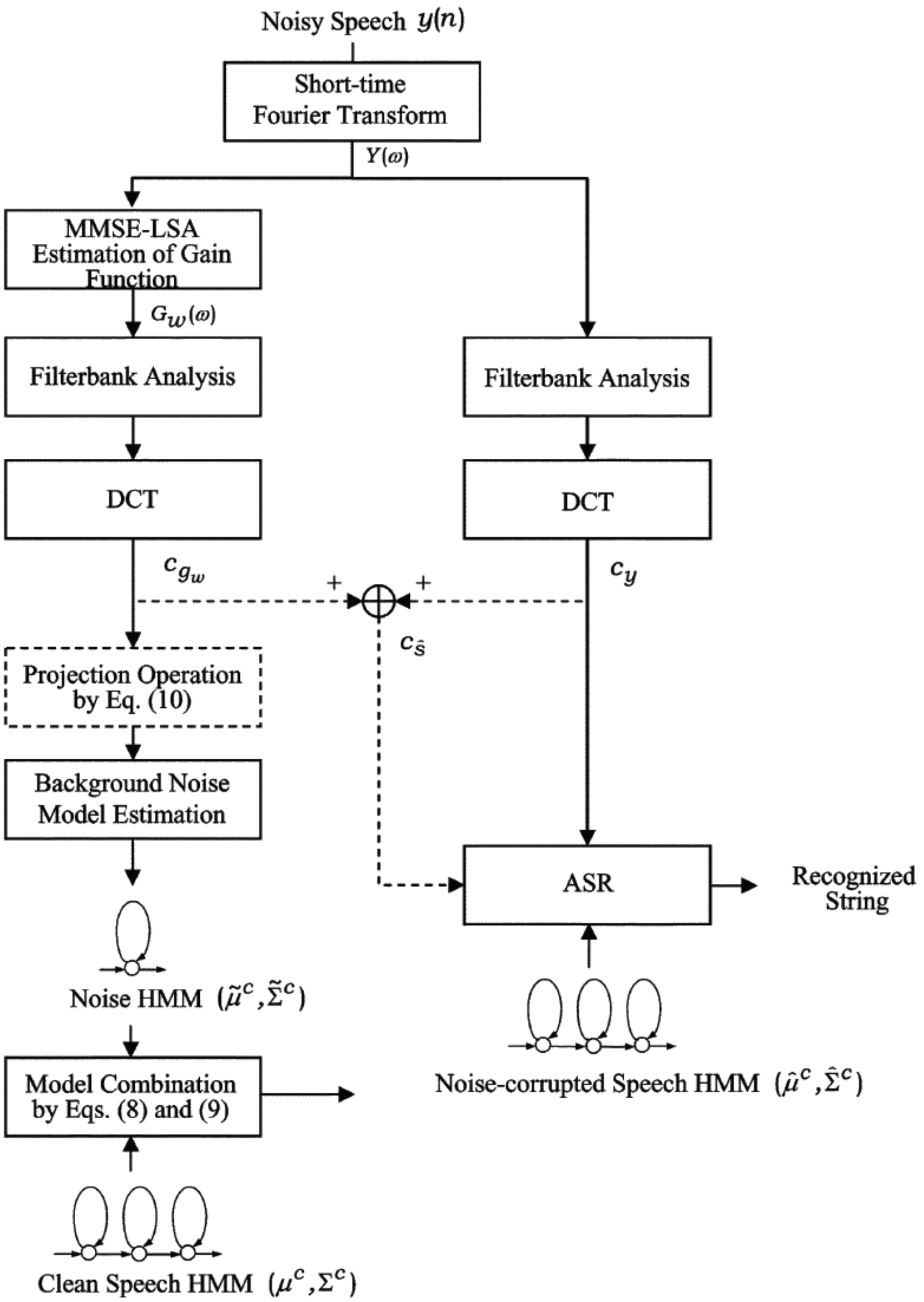
- Introduction
- Additive Model Combination and MMSE-LSA Estimation
- AMC Model Estimation
- Experiments

Abstract—**This paper presents an efficient method for combining models of speech and noise for robust speech recognition applications in noisy environments. This method decomposes the cepstrum domain representation of noise-corrupted speech into clean speech cepstrum and background noise cepstrum components using a minimum mean squared error–log spectral amplitude (MMSE-LSA) criterion. Speech recognition is then performed on noisy cepstrum domain observations using a model that is formed by parallel combination of cepstrum domain clean speech distributions and background noise distributions estimated using this MMSE-LSA based noise decomposition. This method is far more efficient than other parallel model combination (PMC) procedures because model combination is performed directly in the cepstrum domain rather than in the linear spectral domain. Whereas background noise model estimation is addressed as a separate issue in existing PMC procedures, this method explicitly incorporates a mechanism to continually update background noise models and signal-to-noise ratio (SNR) estimates over time. The performance of the proposed cepstrum-domain model combination method is compared with a well known implementation of PMC which uses a log-normal approximation when combining speech and background noise model means and variances on a connected digit string recognition task which is subjected to mismatched channel and environment conditions. As a result, it is shown that the proposed model combination technique gives a word error rate that is comparable to PMC when background noise information and SNR are known prior to estimation. The paper will also present the results of experiments where a combination of cepstrum-domain feature compensation and model combination are applied to this task.**

Introduction

- We focus on the category of acoustic model combination which involves combining acoustic models of speech and noise and performing recognition on the noisy acoustic features. This class of approaches is collectively referred to as parallel model combination (PMC).

- Additive model combination (AMC) is introduced as a special case of PMC where speech and noise are decoupled through MMSE-LSA estimation.


Noisy Speech $y(n)$
Short-time Fourier Transform
$Y(\omega)$
MMSE-LSA Estimation of Gain Function
$G_w(\omega)$
Filterbank Analysis
DCT
c_{g_w}
Projection Operation by Eq. (10)
Background Noise Model Estimation
Noise HMM $(\tilde{\mu}^c, \tilde{\Sigma}^c)$
Model Combination by Eqs. (8) and (9)
Clean Speech HMM (μ^c, Σ^c)
+
+
$c_{\hat{s}}$
Filterbank Analysis
DCT
c_y
ASR
Recognized String
Noise-corrupted Speech HMM $(\hat{\mu}^c, \hat{\Sigma}^c)$

AMC and MMSE-LSA Estimation

- Noisy speech $y(n)$ is represented in the time domain as

  $$y(n) = s(n) + w(n)$$

  - where $s(n)$ and $w(n)$ are clean speech and additive background noise, respectively

- Using a form of **MMSE-LSA estimation**, a gain function $G_w(w)$ can be estimated in the linear frequency domain.

  $$\hat{S}(w) = G_w(w)Y(w)$$

  - where $Y(w) = S(w) + W(w)$ with speech spectrum $S(w)$ and background noise spectrum $W(w)$.

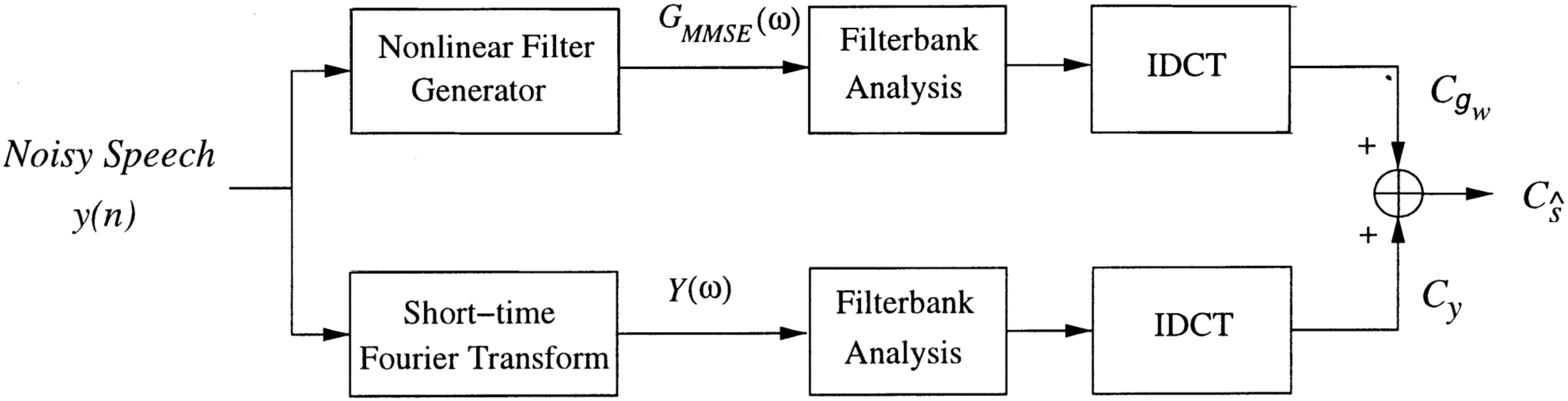
AMC and MMSE-LSA Estimation

- The cepstrum of the gain function c_{gw} represents the background noise term corresponding to the difference between the noise corrupted speech cepstrum c_y and clean speech cepstrum $c_{\hat{s}}$

$$c_{\hat{s}} = c_y + c_{gw}$$

- The cepstrum subtraction method (CSM), obtained an estimate of the *i*th component of the clean speech cepstrum vector at frame τ according to the relationship given in

$$c_{\hat{s},i}(\tau) = c_{y,i}(\tau) + c_{gw,i}(\tau)$$


Noisy Speech
$y(n)$
Nonlinear Filter Generator
$G_{MMSE}(\omega)$
Filterbank Analysis
IDCT
C_{g_w}
+
$C_{\hat{s}}$
+
Short–time Fourier Transform
$Y(\omega)$
Filterbank Analysis
IDCT
C_y

AMC and MMSE-LSA Estimation

- Let $S(w) = A(w)e^{j\rho(w)}$, $W(w)$ and $Y(w) = R(w)e^{j\vartheta(w)}$ be the Fourier expansions of clean speech $s(n)$, additive noise $w(n)$, and noisy speech $y(n)$, respectively.

- The objective of MMSE-LSA is to find the estimator $\hat{A}(w)$ that minimizes the distortion measure $E\{(\log A(w) - \log \hat{A}(w))^2\}$ for the given noisy observation spectrum $Y(w)$.

AMC and MMSE-LSA Estimation

- This gain function is decomposed into two components

$$\hat{A}(w) = G_w(w)R(w) = G_M(w)G_{LSA}(w)R(w)$$

  - where $G_{LSA}(w)$ is the gain function associated with the optimal MMSE-LSA estimator and $G_M(w)$ is a gain modification function.

- The gain $G_{LSA}(w)$ function is $G_{LSA}(\omega) = \frac{\xi(\omega)}{1+\xi(\omega)} \exp(\frac{1}{2} \int_{v(\omega)}^{\infty} \frac{e^{-t}}{t} dt)$

- The gain modification $G_M(\omega)$ function is $G_M(\omega) = \frac{\Lambda(\omega)}{1+\Lambda(\omega)}$

AMC and MMSE-LSA Estimation

- The gain function:

$$G_{LSA}(\omega) = \frac{\xi(\omega)}{1+\xi(\omega)} \exp(\frac{1}{2} \int_{v(\omega)}^{\infty} \frac{e^{-t}}{t} dt)$$

  – where

$$v(\omega) = \frac{\xi(\omega)}{1+\xi(\omega)} \gamma(\omega),\ \gamma(\omega) = \frac{R^2(\omega)}{\lambda_\omega(\omega)}, \xi(\omega) = \frac{\eta(\omega)}{1\text{-}q(\omega)},\ \eta(\omega) = \frac{\lambda_s(\omega)}{\lambda_\omega(\omega)}$$

$$\lambda_s(\omega) = E\{|S(\omega)|^2\} = E\{A^2(\omega)\}$$

$$\lambda_W(\omega) = E\{|W(\omega)|^2\}$$

- The quantity $\gamma(\omega)$ is the estimate of the a posteriori SNR and $\eta(\omega)$ is the estimate of the a priori SNR.

AMC and MMSE-LSA Estimation

- The gain modification function is applied to take into account the probability of the speech presence.

$$G_M(\omega) = \frac{\Lambda(\omega)}{1+\Lambda(\omega)}$$

- $\Lambda(\omega)$ is a likelihood ratio between speech presence and speech absence at frequency ω and is defined by

$$\Lambda(\omega) = \frac{1\text{-}q(\omega)}{q(\omega)} \frac{\exp(v(\omega))}{1+\xi(\omega)}\Bigg|_{\xi(\omega)=\frac{\eta(\omega)}{1\text{-}q(\omega)}}$$

AMC and MMSE-LSA Estimation

- This a priori probability is defined as

$$q(w) = {P(H_o(w))} \Big/ {(P(H_o(w)) + P(H_1(w)))}$$

  - where $P(H_o(w))$ and $P(H_1(w))$ are the probabilities associated with speech absence and speech presence hypotheses, H_o and H_1, respectively.

- In the regions of speech absence:

$$q(w) \to 1 \Rightarrow \Lambda(\omega) \to 0 \Rightarrow G_M(w) \to 0$$

- In the regions of speech presence:

$$q(w) \to 0 \Rightarrow \Lambda(\omega) >> 1 \Rightarrow G_M(w) \to 1$$

AMC Model Estimation

- It is assumed that an HMM model has been trained from uncorrupted utterances. Given this HMM model, the goal is to update the model means and variances to obtain parameters that describe the noise corrupted observations.

$$\hat{u}_i^c = E\{c_{y,i}(\tau)\} = E\{c_{\hat{s},i}(\tau) - c_{gw,i}(\tau)\} = u_i^c - \tilde{u}_i^c$$

$$\begin{aligned}\hat{\sum}_{ij}^c &= E\{(c_{y,i}(\tau) - \hat{u}_i^c)(c_{y,j}(\tau) - \hat{u}_j^c)^T\} \\ &= \sum_{ij}^c + \tilde{\sum}_{ij}^c - C_{ij} - C_{ji}\end{aligned}$$

AMC Model Estimation

$$u_i^c = E\{c_{\hat{s},i}(\tau)\},\ \tilde{u}_i^c = E\{c_{gw,i}(\tau)\}$$

$$\sum_{ij}^c = E\{(c_{\hat{s},i}(\tau) - u_i^c)(c_{\hat{s},i}(\tau) - u_j^c)^T\}$$

$$\tilde{\sum}_{ij}^c = E\{(c_{gw,i}(\tau) - \tilde{u}_i^c)(c_{gw,i}(\tau) - \tilde{u}_j^c)^T\}$$

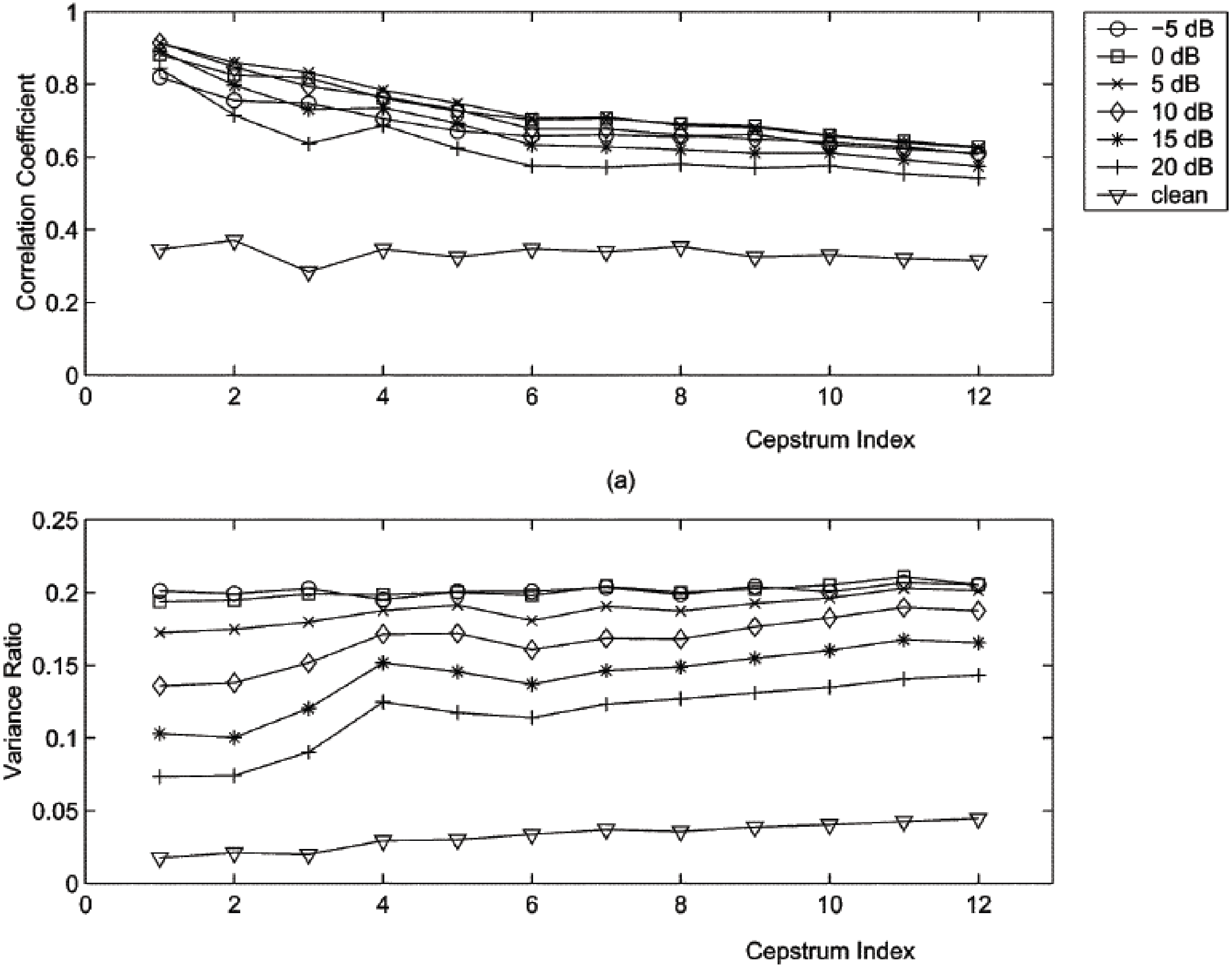
- In addition, C_{ij} and C_{ji} are the covariance matrices between the estimates of background noise cepstrum and clean speech cepstrum.

$$C_{ij} = E\{(c_{\hat{s},i}(\tau)c_{gw,j}(\tau)^T\} - u_i^c\tilde{u}_j^c$$

$$C_{ji} = E\{(c_{\hat{s},j}(\tau)c_{gw,i}(\tau)^T\} - u_j^c\tilde{u}_i^c$$

AMC Model Estimation

- A simple experiment was performed to demonstrate that the noise cepstrum and the estimate of the clean speech cepstrum are highly correlated.
- Except for the clean condition, the correlation between estimated noise cepstrum and clean speech cepstrum is shown in Fig. 2(a) to be high.
- We computed the average ratio of noise cepstrum variance to clean speech cepstrum variance for each SNR condition and displayed it in Fig. 2(b).
- This suggests that the noise cepstrum variance term can be ignored without having significant impact on the updated covariance $\hat{\sum}_{ij}^{c}$.


Correlation Coefficient
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
Cepstrum Index
−5 dB
0 dB
5 dB
10 dB
15 dB
20 dB
clean
Variance Ratio
0
0.05
0.1
0.15
0.2
0.25
0
2
4
6
8
10
12
Cepstrum Index

(a)

(b)

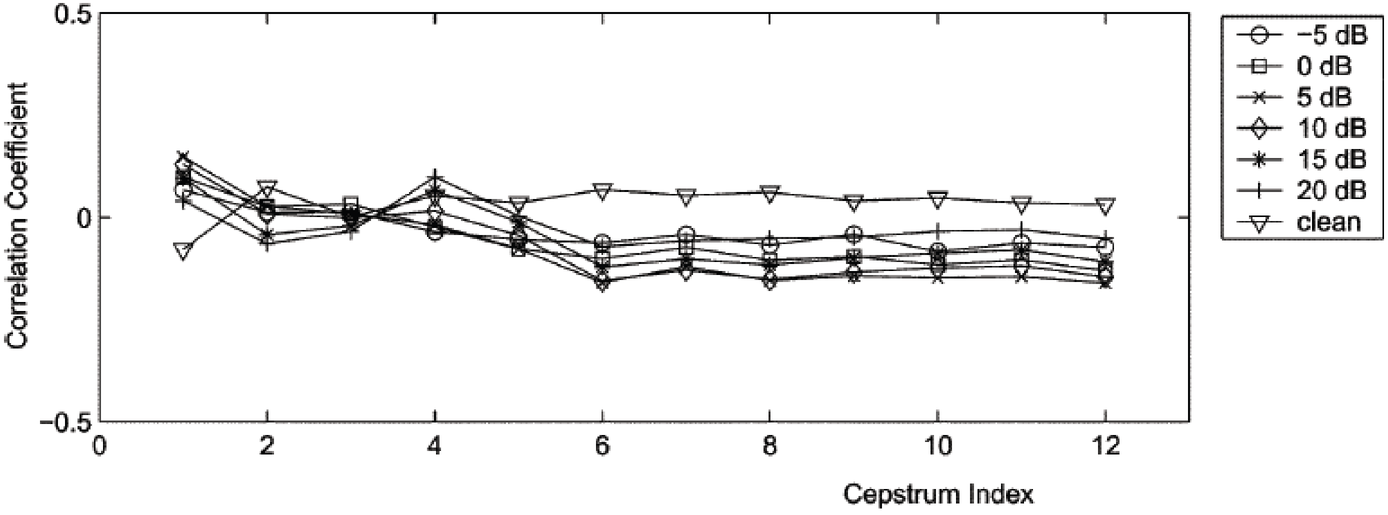
AMC Model Estimation

- The projection operation used is

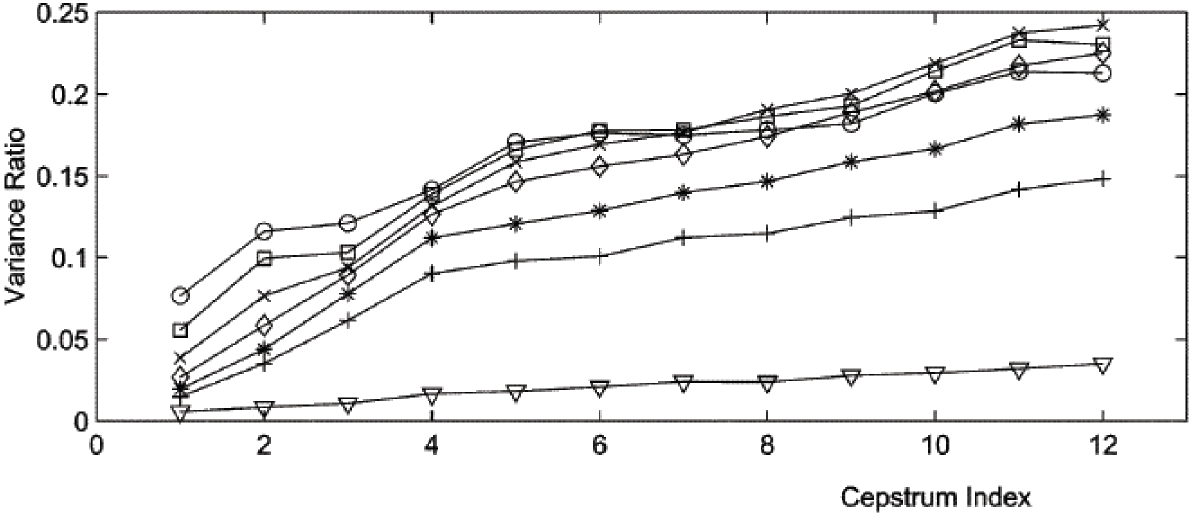
$$\bar{c}_{gw}(\tau) = (c_{gw}(\tau) - \tilde{u}^{c}) - \frac{\left\langle c_{\hat{s}}(\tau) - u^{c}, c_{gw}(\tau) - \tilde{u}^{c} \right\rangle}{\left\langle c_{\hat{s}}(\tau) - u^{c}, c_{\hat{s}}(\tau) - u^{c} \right\rangle}(c_{\hat{s}}(\tau) - u^{c})$$

  - where <x, y> is the inner product of two vectors x and y, and we call $\bar{c}_{gw}(\tau)$ the projected noise cepstrum.

- We will refer to the approach as projected AMC (PAMC).


0.5
0
−0.5
Correlation Coefficient
0
2
4
6
8
10
12
Cepstrum Index
−5 dB
0 dB
5 dB
10 dB
15 dB
20 dB
clean

(a)


0.25
0.2
0.15
0.1
0.05
0
Variance Ratio
0
2
4
6
8
10
12
Cepstrum Index

(b)

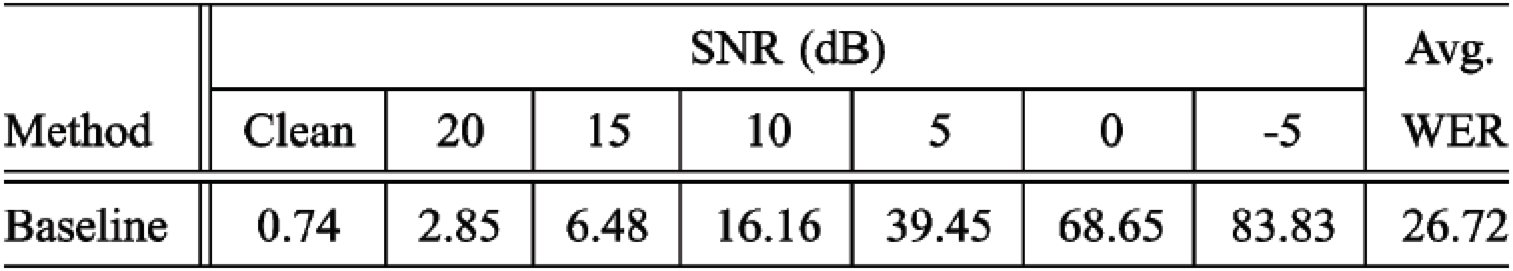
Experiments

- Aurora 2
- we perform ASR experiments in the context of

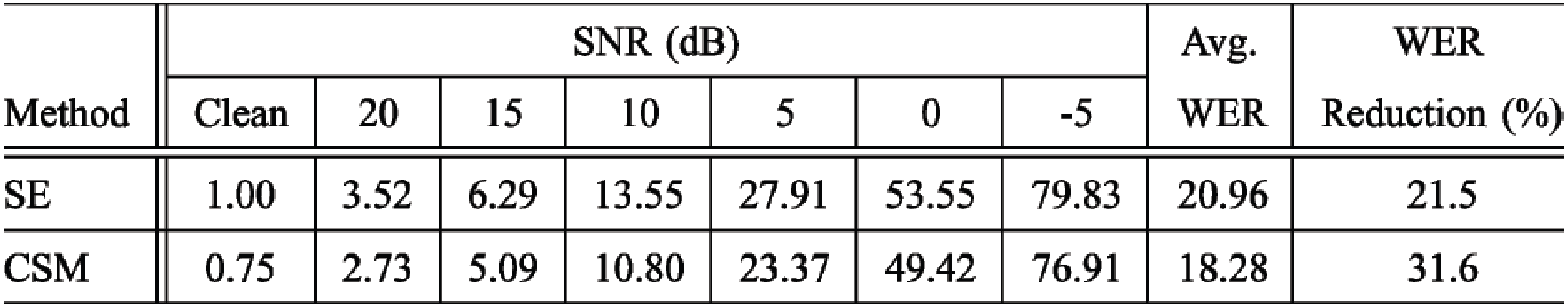
  1) a preprocessing approach that utilizes MMSE-based speech enhancement prior to feature extraction

  2) the acoustic feature compensation or cepstrum subtraction method (CSM) .

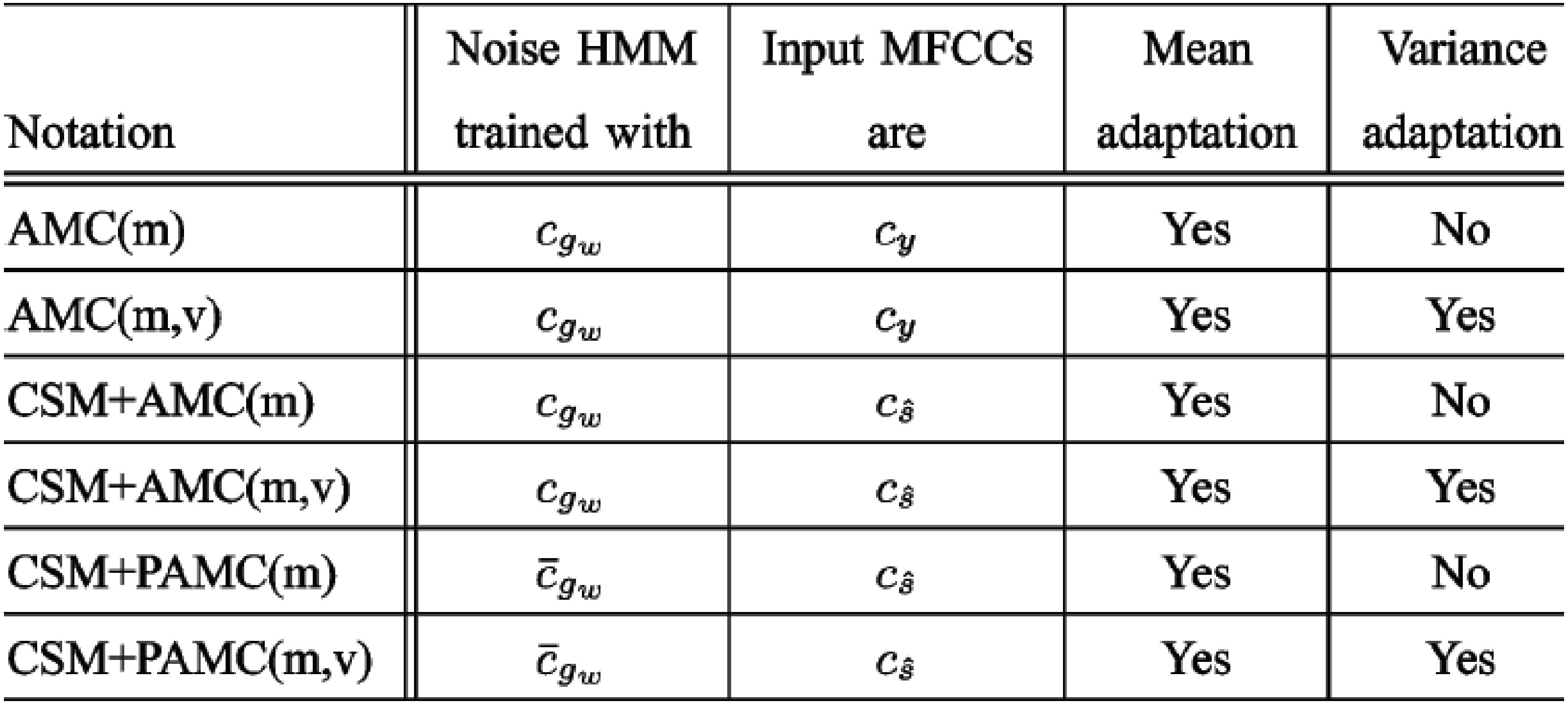
  3)we evaluate the performance of AMC by implementing it using either mean-only adaptation or mean-variance adaptation.



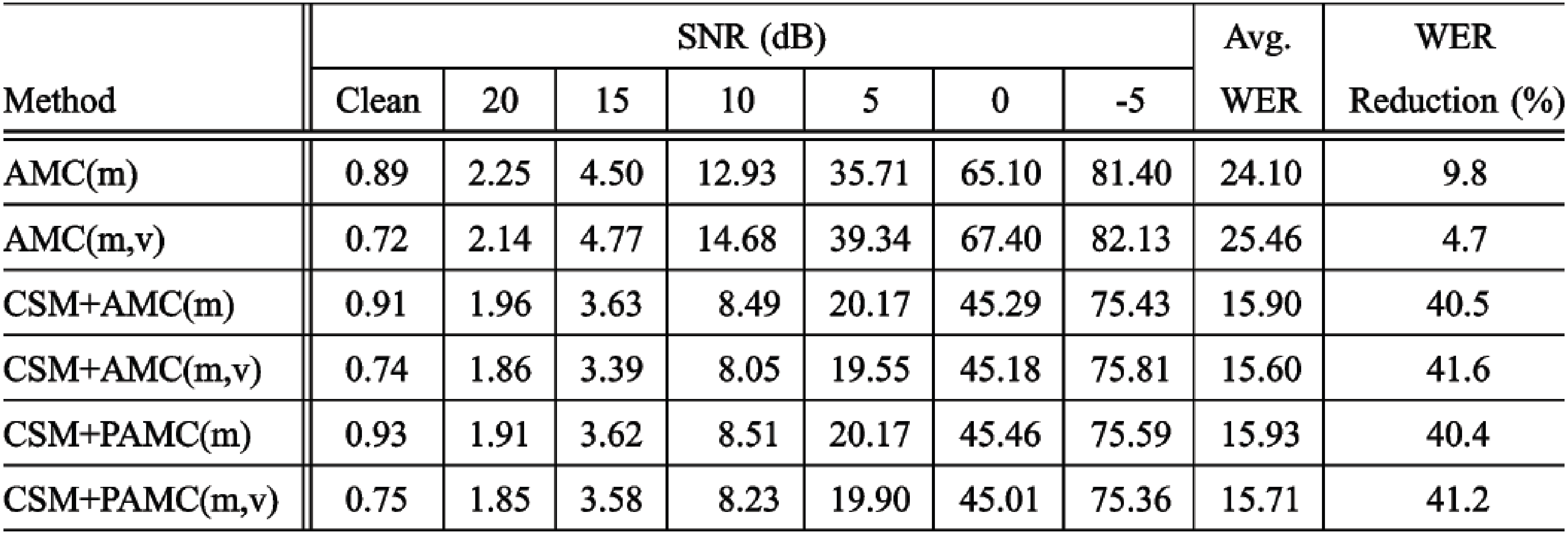
Method	SNR (dB)							Avg.
	Clean	20	15	10	5	0	-5	WER
Baseline	0.74	2.85	6.48	16.16	39.45	68.65	83.83	26.72



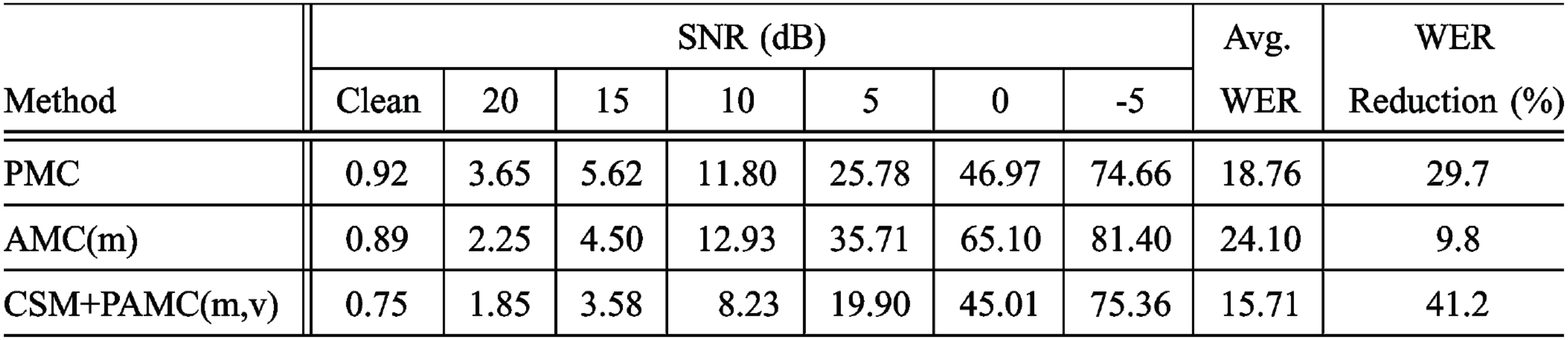
Method	SNR (dB)							Avg.	WER
	Clean	20	15	10	5	0	-5	WER	Reduction (%)
SE	1.00	3.52	6.29	13.55	27.91	53.55	79.83	20.96	21.5
CSM	0.75	2.73	5.09	10.80	23.37	49.42	76.91	18.28	31.6



Notation	Noise HMM trained with	Input MFCCs are	Mean adaptation	Variance adaptation
AMC(m)	c_{g_w}	c_y	Yes	No
AMC(m,v)	c_{g_w}	c_y	Yes	Yes
CSM+AMC(m)	c_{g_w}	$c_{\hat{s}}$	Yes	No
CSM+AMC(m,v)	c_{g_w}	$c_{\hat{s}}$	Yes	Yes
CSM+PAMC(m)	$\bar{c}_{g_w}$	$c_{\hat{s}}$	Yes	No
CSM+PAMC(m,v)	$\bar{c}_{g_w}$	$c_{\hat{s}}$	Yes	Yes



Method	SNR (dB)							Avg.	WER
	Clean	20	15	10	5	0	-5	WER	Reduction (%)
AMC(m)	0.89	2.25	4.50	12.93	35.71	65.10	81.40	24.10	9.8
AMC(m,v)	0.72	2.14	4.77	14.68	39.34	67.40	82.13	25.46	4.7
CSM+AMC(m)	0.91	1.96	3.63	8.49	20.17	45.29	75.43	15.90	40.5
CSM+AMC(m,v)	0.74	1.86	3.39	8.05	19.55	45.18	75.81	15.60	41.6
CSM+PAMC(m)	0.93	1.91	3.62	8.51	20.17	45.46	75.59	15.93	40.4
CSM+PAMC(m,v)	0.75	1.85	3.58	8.23	19.90	45.01	75.36	15.71	41.2



Method	SNR (dB)							Avg.	WER
	Clean	20	15	10	5	0	-5	WER	Reduction (%)
PMC	0.92	3.65	5.62	11.80	25.78	46.97	74.66	18.76	29.7
AMC(m)	0.89	2.25	4.50	12.93	35.71	65.10	81.40	24.10	9.8
CSM+PAMC(m,v)	0.75	1.85	3.58	8.23	19.90	45.01	75.36	15.71	41.2