Static and Dynamic Variance Compensation for Recognition of Reverberant Speech With Dereverberation Preprocessing

Author : Marc Delcroix, Tomohiro Nakatani and Shinji Watanabe

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Dereverberation Based on Late Reverberation Energy Suppression
- Proposed Method for Variance Compensation
- Experiments

Introduction

- The performance of automatic speech recognition is severely degraded in the presence of **noise or reverberation**.

- In this paper, we use a **dereverberation method** to reduce reverberation prior to recognition.

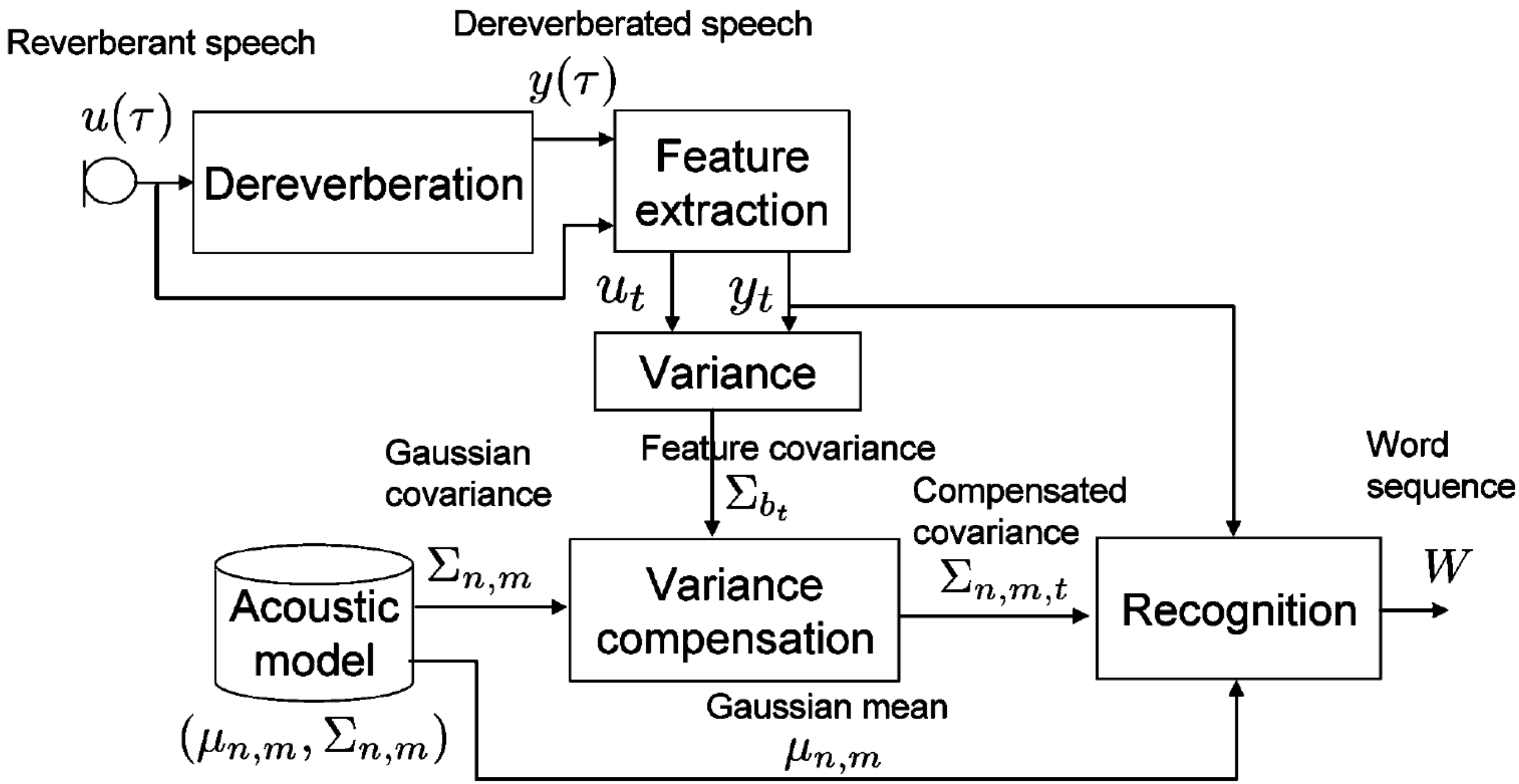

Fig. 1. Schematic diagram of recognition system for reverberant speech.

Dereverberation

- Reverberant speech $u(\tau)$ is usually modeled as the convolution of clean speech $x(\tau)$ with a room impulse response $h(\tau)$ as

$$u(\tau) = x(\tau) * h(\tau)$$

- Let us divide the room impulse response into two parts : early reflections and late reflections.

$$u(\tau) = x(\tau) * h_c(\tau) + x(\tau) * h_l(\tau)$$

Dereverberation

- In this paper, we use a dereverberation method that focuses on late reverberation removal.

- We can show that an **estimate of the late reverberation** $l(\tau)$ can be approximated by a convolution of observed reverberant speech $u(\tau)$ with a linear prediction filter $w_D(\tau)$ as

$$l(\tau) = w_D(\tau) * u(\tau)$$

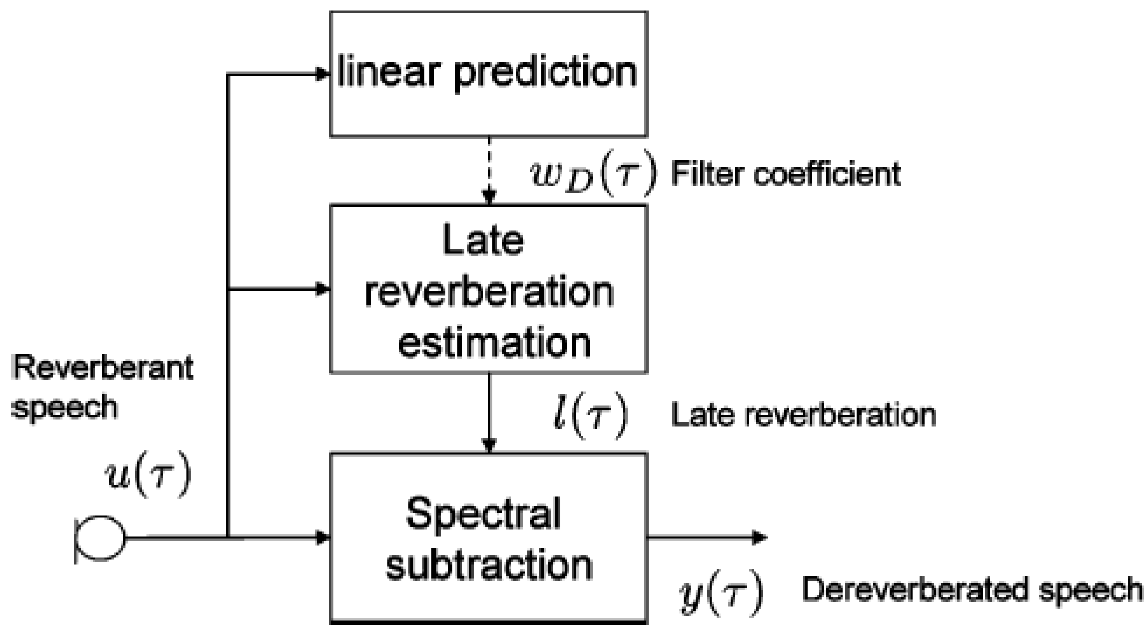

Fig. 2. Schematic diagram of the dereverberation preprocessor.

Variance Compensation

- Speech is modeled using a hidden Markov model (HMM) with the state density modeled by a Gaussian mixture (GM)

$$
\begin{aligned}
p(x_t|n) &= \sum_{m=1}^{M} p(m)p(x_t|n,m) \\
&= \sum_{m=1}^{M} p(m)N(x_t;\mu_{n,m},\Sigma_{n,m})
\end{aligned}
$$

n: state index

m: the Gaussian mixture component index

$\mu_{n,m}$:mean vector

$\Sigma_{n,m}$: covariance matrix

Variance Compensation

- Let us model the mismatch b_t between clean speech and reverberant speech as

  $$u_t = x_t + b_t$$

  where u_t is the observed reverberant speech feature and b_t is modeled as a Gaussian with

  $$p(b_t) \approx N(b_t; \hat{b}_t, \sum b_t)$$

  where $\hat{b}_t$ is an estimate of the mismatch, i.e., $\hat{b}_t = u_t - y_t$, y_t is the dereverberated speech feature, and $\sum b_t$ represents a time-varying feature covariance matrix.

Variance Compensation

- The likelihood of a reverberant speech feature given a state n can be obtained by marginalizing the joint probability over clean speech

$$
\begin{aligned}
p(u_t|n) &= \int_{-\infty}^{+\infty} p(u_t, x_t|n)dx_t \\
&= \int_{-\infty}^{+\infty} p(u_t|x_t, n)p(x_t|n)dx_t
\end{aligned}
$$

Variance Compensation

- It is assumed that $p(u_t|x_t, n) = p(b_t|x_t, n) \approx p(b_t)$

$$p(u_t|n) = \int_{-\infty}^{+\infty} \sum_{m=1}^{M} p(m) N(x_t, \mu_{n,m}, \Sigma_{n,m})$$
$$N(u_t - x_t; u_t - y_t, \mathbf{\Sigma}_{b_t}) dx_t$$
$$= \sum_{m=1}^{M} p(m) N(y_t; \mu_{n,m}, \underbrace{\mathbf{\Sigma}_{n,m} + \mathbf{\Sigma}_{b_t}}_{\triangleq \mathbf{\Sigma}_{n,m,t}})$$

Variance Compensation

- The compensated mixture covariance matrix is modeled as

  $$\sum_{n,m,t}^{'} = \sum_S + \sum_D$$

  where $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_D$ represent **static** and **dynamic** variance components, respectively.

- We further express $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_D$ with a parametric representation similar to MLLR. The **static variance** $\mathbf{\Sigma}_S$ can thus be expressed as

  $$\sum_S (\sum_{n,m}, L) = L \sum_{n,m} L^T$$

  where $\mathbf{L}$ is a matrix of static variance compensation parameters.

Variance Compensation

- If we assume the use of a **diagonal covariance matrix**, which is widely employed in speech recognition
  $$(\Sigma_S(\Sigma_{n,m},\lambda))_{i,i} = \lambda_i \sigma^2_{n,m,i}$$
  where λ_i can be interpreted as the weight of the variances of the acoustic models.

- In a similar way, we model the **dynamic variance** as
  $$\Sigma_D(\Sigma_{b_t}, A) = A\Sigma_{b_t} A^T$$
  where A is a matrix of dynamic variance compensation parameters.

Variance Compensation

- We can express the **dynamic variance** component as
  $$(\Sigma_D(\hat{b}_t, \alpha))_{i,i} = \alpha_i \hat{b}_{t,i}^2$$
  where α_i are model parameters.

- Therefore, with the proposed model, we can rewrite the time-varying state variance as
  $$(\Sigma'_{n,m,t})_{i,i} = \alpha_i \hat{b}_{t,i}^2 + \lambda_i \sigma_{n,m,i}^2$$

Variance Compensation

- The model variance parameters, $\theta = (\alpha, \lambda)$, can be obtained by maximizing the likelihood as

$$(\theta, W) = \arg\max_{\theta, W} p(U \mid W, \theta) p(W)$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T]$ is a sequence of observed speech features. The word sequence W is known.

Variance Compensation

- The maximum-likelihood estimation problem can be solved using the EM algorithm. We define an auxiliary function $Q(\theta|\bar{\theta})$ as

$$Q(\theta|\bar{\theta}) = \sum_{S}\sum_{C}\int\int_{X+B=U} p(X,B,S,C|\Psi,\bar{\theta}) \times \log(p(X,B,S,C|\Psi,\theta))dXdB$$

  Where *B* is a mismatch feature sequence, *S* is a set of all possible state sequences, *C is* a set of all mixture components, Ψ represents the acoustic model parameters, and $\bar{\theta}$ represents an estimate of parameter θ obtained from the previous step of the EM algorithm.

Variance Compensation

$$Q(\theta \mid \bar{\theta}) \propto \sum_{S} \sum_{C} \iint_{X+B=U} p(X, B, S, C \mid \psi, \bar{\theta})$$

$$\times \log\left(\left(\prod_{t=1}^{T} p(b_t \mid \alpha)\, p(x_t \mid s_t = n, c_t = m, \lambda)\right) \right) dX dB$$

$$= \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \iint_{X+B=U} p(X, B, n, m \mid \psi, \bar{\theta}) \log(p(b_t \mid \alpha)) dX dB$$

$$+ \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \iint_{X+B=U} p(X, B, n, m \mid \psi, \bar{\theta}) \log(p(x_t \mid n, m, \lambda)) dX dB$$

Variance Compensation

$$\lambda_i = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\frac{R(x_{t,i},y_t,n,m,\Psi,\bar{\alpha},\bar{\lambda})}{\bar{\lambda}_i\sigma^2_{n,m,i}}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)}$$

- $R(x_{t,i}, y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda})$ is an estimate of the dereverberated feature variance. It is given by

$$\begin{aligned} &R(x_{t,i}, y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}) \\ &\quad = \mu^2_{n,m,i} - 2\mu_{n,m,i}E\{x_{t,i}|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\} \\ &\quad + E\{x^2_{t,i}|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\} \end{aligned}$$

Variance Compensation

- $E\{x_{t,i}|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}$ is an estimate of the clean speech feature, expressed as

$$E\{x_{t,i}|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}$$
$$= \frac{\int\int_{X+B=U} x_{t,i} p(x_t, b_t, n, m|\Psi, \bar{\theta}) dx_t db_t}{p(y_t|n, m, \Psi, \bar{\theta})}$$
$$= \frac{\bar{\alpha}_i \hat{b}_{t,i}^2 \bar{\lambda}_i \sigma_{n,m,i}^2}{\bar{\alpha}_i \hat{b}_{t,i}^2 + \bar{\lambda}_i \sigma_{n,m,i}^2} \left(\frac{y_{t,i}}{\bar{\alpha}_i \hat{b}_{t,i}^2} + \frac{\mu_{n,m,i}}{\bar{\lambda}_i \sigma_{n,m,i}^2} \right)$$

- $E\{x_{t,i}^2|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}$ is an estimate of the clean feature variance, expressed as

$$E\{x_{t,i}^2|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}$$
$$= \frac{\int\int_{X+B=U} x_{t,i}^2 p(x_t, b_t, n, m|\Psi, \bar{\theta}) dx_t db_t}{p(y_t|n, m, \Psi, \bar{\theta})}$$
$$= \frac{\bar{\alpha}_i \hat{b}_{t,i}^2 \bar{\lambda}_i \sigma_{n,m,i}^2}{\bar{\alpha}_i \hat{b}_{t,i}^2 + \bar{\lambda}_i \sigma_{n,m,i}^2} + E\{x_{t,i}|y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}^2$$

$$\int x_{t,i}\, p(x_t \mid y_t, n, m, \psi, \bar{\theta}) dx_t$$

$$= \int x_{t,i} \frac{p(x_t, y_t \mid n, m, \psi, \bar{\theta})}{p(y_t \mid n, m, \psi, \bar{\theta})} dx_t$$

$$= \frac{\iint p(x_t, y_t \mid n, m, \psi, \bar{\theta}) dx_t db_t}{p(y_t \mid n, m, \psi, \bar{\theta})}$$

Variance Compensation

- The counting function: $n_t = (X, B, m, n) = \begin{cases} 1, if \ \mathrm{s} = n, c = m \\ 0, \text{otherwise} \end{cases}$

$$Q(\theta \mid \overline{\theta}) \propto \sum_S \sum_C \iint_{X+B=U} p(X, B, S, C \mid \psi, \overline{\theta})$$
$$\times \log\left((\prod_{t=1}^{T} p(b_t \mid \alpha) p(x_t \mid s_t = n, c_t = m, \lambda) \right) dXdB$$
$$= \sum_S \sum_C \iint_{X+B=U} n_t(X, B, C, s) p(X, B, S, C \mid \psi, \overline{\theta})$$
$$\times \log\left((\prod_{t=1}^{T} p(b_t \mid \alpha) p(x_t \mid s_t = n, c_t = m, \lambda) \right) dXdB$$

$$r_t(n, m) = \sum_S \sum_C (X, B, n, m) p(X, B, S, C \mid \psi, \overline{\theta})$$

Variance Compensation

$$\alpha_i = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_t(n,m) \frac{E\{b_{t,i}^2 | y_t, n, m, \Psi, \bar{\alpha}, \bar{\lambda}\}}{\hat{b}_{t,i}^2}}{\sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_t(n,m)}$$

Variance Compensation

$$E\left\{b_{t,i}^2|y_t,n,m,\Psi,\bar{\alpha},\bar{\lambda}\right\}$$

$$= \frac{\int\int_{X+B=U} b_{t,i}^2 p(x_t,b_t,n,m|\Psi,\bar{\theta})dx_t db_t}{p(y_t|n,m,\Psi,\bar{\theta})}$$

$$= \frac{\bar{\alpha}_i \hat{b}_{t,i}^2 \bar{\lambda}_i \sigma_{n,m,i}^2}{\bar{\alpha}_i \hat{b}_{t,i}^2 + \bar{\lambda}_i \sigma_{n,m,i}^2} + E\{b_{t,i}|y_t,n,m,\Psi,\bar{\alpha},\bar{\lambda}\}^2$$

$$E\left\{b_{t,i}|y_t,n,m,\Psi,\bar{\alpha},\bar{\lambda}\right\}$$

$$= \frac{\int\int_{X+B=U} b_{t,i} p(x_t,b_t,n,m|\Psi,\bar{\theta})dx_t db_t}{p(y_t|n,m,\Psi,\bar{\theta})}$$

$$= \frac{\bar{\alpha}_i \hat{b}_{t,i}^2}{\bar{\alpha}_i \hat{b}_{t,i}^2 + \bar{\lambda}_i \sigma_{n,m,i}^2}(y_{t,i} - \mu_{n,m,i}).$$

Experiments

- NTT Corporation
- The recognition task consisted of continuous digit utterances.
- The acoustic features consisted of 39 coefficients: 12 MFCCs, the 0th cepstrum coefficient, delta, and acceleration.

	WER (%)
Clean	1.2
Reverberant	32.7
Dereverberated	31.0
Variance compensation ($\alpha = 1$, $\lambda = 1$)	15.9
Variance compensation	3.3

	1.5 m			2 m		
Reverberant	32.7 %			37.1 %		
Dereverberated	31.0 %			36.3 %		
Variance Compensation (without adaptation)	15.9 %			19.5 %		
	2 ut.	32 ut.	512 ut.	2 ut.	32 ut.	512 ut.
SVA	15.1 %	15.1 %	15.2 %	18.4 %	18.4 %	18.5 %
DVA	15.6 %	15.5 %	15.5 %	19.6 %	19.4 %	19.2 %
SDVA	**13.5 %**	13.3 %	13.4 %	**16.6 %**	16.3 %	16.3 %