HMM-based noise-robust feature compensation

Author : **Akira Sasou, Kazuyo Tanaka**

Professor : 陳嘉平

Reporter : 楊治鏞

Outline

- Introduction
- Feature compensation based on HMM
- Experimental result

Introduction

- A distortion of the noisy speech feature in the cepstral domain, which is different from the clean speech feature, can be divided into a stationary distortion component and a non-stationary distortion component.

- The proposed method eliminates the degradation of feature-compensation accuracy caused by the non-stationary distortion component.

Feature compensation based on HMM

- The filter bank energy (FBE) of the noisy speech x_b can be represented as a function of the clean speech s_b and the noise n_b.

$$x_b = s_b + n_b$$

$$x_l = \log(x_b) = s_l + \log[1 + \exp(n_l - s_l)]$$

- Where the subscript l denotes that the expression is in the log-FBE domain.

Cepstral domain

- The noisy speech x in the cepstral domain can be represented by:

$$x = \boldsymbol{C} \cdot x_l = s + g(s,n)$$

- Where $g(s,n)$ represents a distortion of the noisy speech in the cepstral domain.

$$g(s,n) = \boldsymbol{C} \cdot \log[1 + \exp\{\boldsymbol{C}^{-1} \cdot (n - s)\}]$$

- Here $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ denote the discrete cosine transform (DCT) matrix and its inverse transform matrix.

HMM

- The output pdf of the j th state $b_j(s)$ is given by:

$$b_j(s) = \sum_{m=1}^{M} w_{jm} N(s; \mu_{jm}, \Sigma_{jm})$$

- where M is the number of Gaussian distributions and Σ_{jm} is a diagonal matrix.

Element vector

- Let $\hat{x}$ denote an element vector consisting of the part of an observed feature vector x that is to be compensated.

- In our settings, the vector x has a 39-dimensional vector, which consists of a 13-dimensional base mel-frequency cepstral coefficient (MFCC) and its delta and delta–delta components, and the element vector $\hat{x}$ corresponds to the base MFCC.

Gaussian distribution

- In addition, let $N(\hat{\mu}_{jm}, \hat{\Sigma}_{jm})$ represent the Gaussian distribution of the clean speech $\hat{s}$ corresponding to the element vector $\hat{x}$ of the noisy speech.

Gaussian distribution

- Every noise-adapted Gaussian distribution $N(\hat{\boldsymbol{m}}_{jm}, \hat{\boldsymbol{Z}}_{jm})$ is then given by:

$$\hat{\boldsymbol{m}}_{jm} = E[\hat{\boldsymbol{x}}]_{\hat{\boldsymbol{s}},\hat{\boldsymbol{n}}}$$

$$\hat{\boldsymbol{Z}}_{jm} = \hat{\Sigma}_{jm}$$

- Where $E[\hat{\boldsymbol{x}}]_{\hat{\boldsymbol{s}},\hat{\boldsymbol{n}}}$ represents the expectation of $\hat{x}$ with regard to both $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{n}}$. The $\hat{\boldsymbol{n}}$ is the noise vector corresponding to the element vector $\hat{x}$ of the noisy speech.

Gaussian distribution

$$E[\hat{\boldsymbol{x}}]_{\hat{\boldsymbol{s}}} = \hat{\mu}_{jm} + g(\hat{\mu}_{jm}, \hat{\boldsymbol{n}})$$

- The actual distribution of the noise $\hat{\boldsymbol{n}}$ is not known. Thus, we evaluate the expectation of $E[\hat{\boldsymbol{x}}]_{\hat{\boldsymbol{s}}}$ with regard to $\hat{\boldsymbol{n}}$ as follows:

$$\hat{\boldsymbol{m}}_{jm} = E[\hat{\mu}_{jm} + g(\hat{\mu}_{jm}, \hat{\boldsymbol{n}})]_{\hat{\boldsymbol{n}}} = \hat{\mu}_{jm} + \hat{\boldsymbol{d}}_{jm}$$

$$\hat{\boldsymbol{d}}_{jm} = E[g(\hat{\mu}_{jm}, \hat{\boldsymbol{n}})]_{\hat{\boldsymbol{n}}} \approx \frac{1}{N}\boldsymbol{C} \cdot \sum_{\boldsymbol{t=1}}^{\boldsymbol{N}} \log[1 + \exp\{\boldsymbol{C}^{-1} \cdot (\hat{\boldsymbol{n}}_t - \hat{\mu}_{jm})\}]$$

- Where $\hat{\boldsymbol{n}}_t, (t = 1, \ldots, N)$, are noise vectors.

Forward path probability

- The proposed method utilizes each forward path probability $\alpha(s,t)$ of the best path to the sth state at time t for the feature-compensation process.

$$\alpha(s,0) = \log(\pi_s) \qquad \forall s \in S$$

$$\alpha(s,t) = \max_{j \in S} \alpha(j,t-1) + \log\{a_{js} \cdot b_j(\boldsymbol{y}_t)\}$$

- S denotes the entire set of HMM states, π_s denotes the initial probability of the sth state, and a_{js} denotes the transition probability.
- where the vector $\boldsymbol{y}_t$ represents the compensated feature vector.

Forward path probability

- Using the previously calculated forward path probabilities, we evaluate weights $\alpha'(s,t-1)$ to be applied to the noise-adapted output pdfs of all the states.

$$\alpha'(s,t-1) = \exp\{\frac{\alpha(s,t-1)}{t}\}$$

Posterior probability

- We adopted the following approximation of posterior probability $P(j,m)$ of each noise-adapted Gaussian distribution $N(\hat{\boldsymbol{m}}_{jm}, \hat{\boldsymbol{Z}}_{jm})$:

$$P(j,m) = \frac{\alpha'(j,t-1) w_{jm} N(\hat{\boldsymbol{x}}_t; \hat{\boldsymbol{m}}_{jm}, \hat{\boldsymbol{Z}}_{jm})}{\sum_{s \in S} \sum_{n=1}^{M} \alpha'(s,t-1) w_{sn} N(\hat{\boldsymbol{x}}_t; \hat{\boldsymbol{m}}_{sn}, \hat{\boldsymbol{Z}}_{sn})}$$

- The compensated element vector $\hat{\boldsymbol{y}}_t$ is given by:

$$\hat{\boldsymbol{y}}_t = \hat{\boldsymbol{x}}_t - \sum_{j \in S} \sum_{m=1}^{M} P(j,m) \hat{\boldsymbol{d}}_{jm}$$

Stationary component

- When the noise has no stationary component, such as transient pulse noises, the stationary distortion components are ideally equal to zero: $\hat{\boldsymbol{d}}_{jm} = 0$

- In this case, we need to find another feature-compensation formula without using $\hat{\boldsymbol{d}}_{jm}$.

Feature-compensation formula

$$\hat{\boldsymbol{y}}_t = \sum_{j \in S} \sum_{m=1}^{M} P(j,m)\{\hat{\boldsymbol{x}}_t - \hat{\boldsymbol{d}}_{jm}\}$$

- where $(\hat{x}_t - \hat{\boldsymbol{d}}_{jm})$ can be regarded as an estimate of the clean speech feature assuming that the clean speech feature was emitted from the m th Gaussian distribution in the j th state's pdf.

Feature-compensation formula

- The expectation vector of the Gaussian distribution can be also regarded as an estimate of the clean speech feature, because the expectation vector is the most probable one in the distribution.

$$\hat{\boldsymbol{y}}_t = \sum_{j \in S} \sum_{m=1}^{M} P(j,m) \hat{\mu}_{jm}$$

- The compensated feature vector $\boldsymbol{y}_t$ is finally obtained by combining the compensated element vector $\hat{\boldsymbol{y}}_t$ with the elements of the observed feature vector x_t except the element vector $\hat{\boldsymbol{x}}_t$.

Experimental set-up

- AURORA2
- Method 1
- Method 2

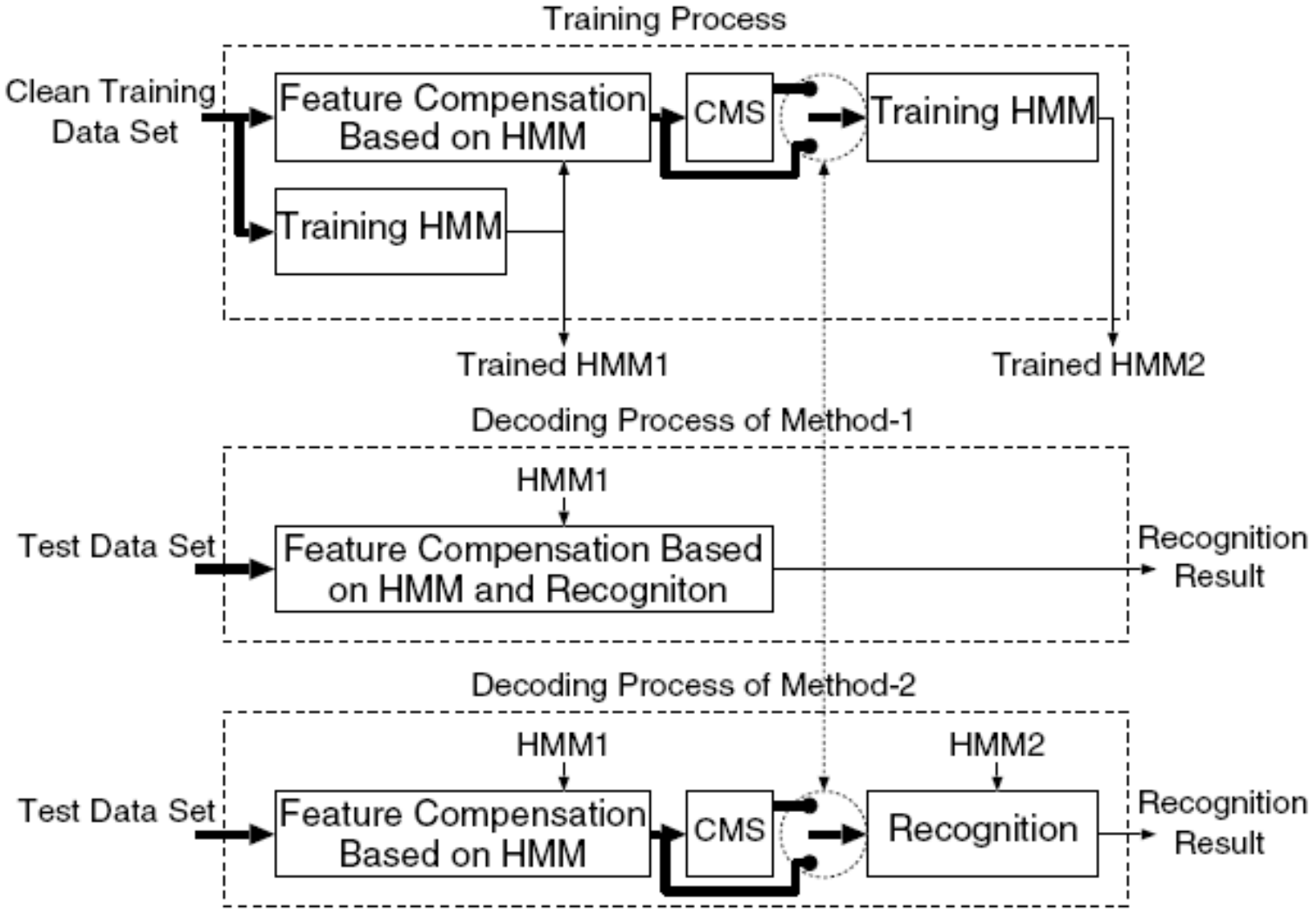


Fig. 1. Flow diagram of the training process and decoding processes for Methods 1 and 2.

Table 4
Relative improvement for Methods 1 and 2

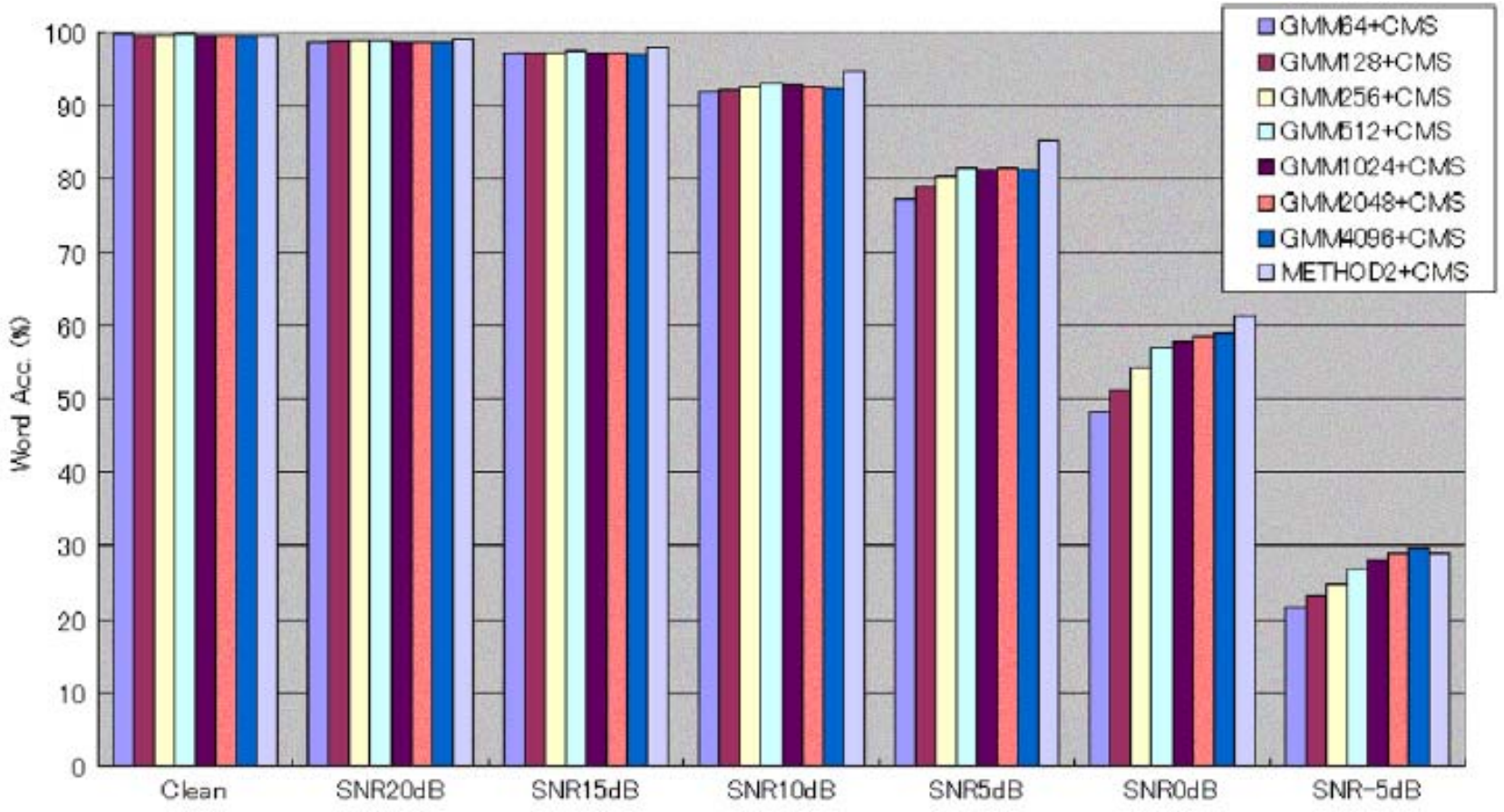
Method	Set A	Set B	Set C	Overall
Method 1	52.18	56.11	37.84	51.84
Method 2	60.98	61.34	48.24	59.25

Table 6
Relative improvement for Method 2 followed by CMS

Method	Set A	Set B	Set C	Overall
Method 2 + CMS	66.18	70.08	54.50	66.24

Comparison of GMM-based method

- To compare the proposed method with the GMM-based feature-compensation method, we conducted experiments using GMM followed by CMS as front-end feature compensation.
- The number of Gaussian distributions ranged from 64 to 4096.
- Each back-end HMM was trained from the compensated features for each condition.


GMM64+CMS
GMM128+CMS
GMM256+CMS
GMM512+CMS
GMM1024+CMS
GMM2048+CMS
GMM4096+CMS
METHOD2+CMS
Word Acc. (%)
100
90
80
70
60
50
40
30
20
10
0
Clean
SNR20dB
SNR15dB
SNR10dB
SNR5dB
SNR0dB
SNR-5dB

Evaluation in a transient pulse noise environment

- In the experiment conducted in a transient pulse noise environment, we used wooden collision sound sources recorded in the Real World Computing Partnership (RWCP) Sound Scene Database.

- We generated noise-corrupted speech data by adding a transient pulse every 125 ms to clean speech data from AURORA2.

Word accuracy evaluated in a transient pulse noise environment

	Baseline	Method 1
Clean	99.66	99.60
20 dB	98.77	98.96
15 dB	98.07	98.46
10 dB	94.44	98.07
5 dB	87.38	96.41
0 dB	75.10	94.23
−5 dB	53.95	89.38
Average	77.26	93.95