IMPROVED MODEL PARAMETER COMPENSATION METHODS FOR NOISE-ROBUST SPEECH RECOGNITION

Y.H.Chang , Y.J.Chung, S.H.Park

LGIC R&D Center
533, Hogye-dong, Dongan-gu, Anyang-shi, Kyongki-do, 431-080, Korea

ABSTRACT

In this paper we study model parameter compensation methods for noise-robust speech recognition based on CDHMM. First, we propose a modified PMC method where adjustment term in the model parameter adaptation is varied depending on mixture components of HMM to obtain more reliable modeling. A state-dependent association factor that controls the average parameter variability of Gaussian mixtures and the variability of the respective mixtures is used to find the final optimum model parameters. Second, we propose a re-estimation solution of environmental variables with additive noise and spectral tilt based on expectation-maximization (EM) algorithm in the cepstral domain. The approach is based on the vector Taylor series (VTS) approximation. In our experiments on a speaker independent isolated Korean word recognition, the modified PMC show better performance than the Gales' PMC and the proposed VTS is superior to both of them.

1. INTRODUCTION

HMM-based speech recognition systems have been widely used to reduce recognition error rates in adverse environments. A model parameter compensation method is one of the noise-robust speech recognition methods. This method reduces many computational complexities while feature vector transformation or noise robust auditory modeling methods have computational loads in retraining the recognizer. Also, we can perform model parameter compensation by using only the testing words, even in case that no stereo databases exist. In addition, model parameter compensation methods have been fast developed in various approaches.

In this paper we use only mean vectors of the Gaussian mixtures, and the mean vectors are compensated by estimating noise mean vectors state by state. The method of adjusting only mean vectors has an advantage in reducing computational times and prevents us from using incorrect noise covariance estimated from 3 or 4 frames of noise.

This paper is organized as follows. In section 2, we discuss the proposed model parameter compensation algorithms, so called EM-driven and steepest-descent-driven PMC with an association factor and mean-VTS0. EM-driven PMC and steepest-descent-driven PMC adapt clean speech mean vectors to noisy speech with an association factor calculated by the EM and the steepest descent algorithm, respectively. Mean-VTS0 compensates clean speech mean vectors by using spectral tilt vectors as well as noise mean vectors estimated by the 0th order vector Taylor series approximation. In section 3, we discuss our experimental results on speaker-independent Korean isolated word recognition corrupted by additive white Gaussian noise and driving car noise. Finally, we summarize our outcomes and discuss our future work.

2. PROPOSED MODEL COMPENSATION ALGORITHMS

2.1. EM-driven PMC and steepest-descent-driven PMC with an association factor

In this sub-section we discuss two new methods of model compensation that have high recognition accuracy in low SNR. These algorithms are based on PMC, but are very simple. The algorithms use noise and clean speech model parameters, and noise model parameters are estimated from 3 or 4 frames at the beginning of noisy speech. The PMC method proposed by Gales and Young[1][2] uniformly compensates all clean speech model parameters such as mean and covariance of the Gaussian mixtures with estimated noise model parameters. In spite of the fact that a recognizer based on HMM estimates model parameters in every state, the PMC method proposed by Gales and Young adjusts those without considering individual information, such as the different effect of the uniform compensation between states, mixtures in a state. In this paper, we propose new model parameter compensation algorithms that are based on the PMC. They are called respectively EM-driven and steepest-descent-driven PMC with an association factor. Assumptions for the new algorithms are as follows.

- Variation characteristic of model parameters is similar in a state, but is different between states.
- Covariance of Gaussian mixtures is invariable. So a covariance matrix of a noisy speech is equal to that of the clean speech.
- All of the mean vectors of the Gaussian mixtures are adapted state by state.

The above assumptions are expressed as follows within a state.

$$\begin{aligned} \mathbf{b}_{1i} &= \hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i, \quad i = 0,1,\cdots,M-1 \\ \mathbf{b}_2 &= \frac{1}{M}\sum_{i=0}^{M-1}(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) = E\{\mathbf{b}_1\} \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\boldsymbol{\mu}_i$ are separately a mean vector compensated by the Gales' PMC and a mean vector of clean speech in i-th Gaussian mixture. The vector $\mathbf{b}_{1i}$ is an amount of difference of a mean vector of i-th Gaussian mixture. The constant M is a

number of Gaussian mixtures in a state, and the vector $\mathbf{b}_2$ expresses an average variation of mean vectors in a state.

We can evaluate a vector $\mathbf{b}$ called the bias vector in each state that is used for compensating differences of mean vectors between clean speech and noisy speech. Then, the bias vector $\mathbf{b}$ is

$$\mathbf{b} = (1-\lambda)\mathbf{b}_{1i} + \lambda\mathbf{b}_2 \quad (2)$$

where the unknown constant λ weights between the vector $\mathbf{b}_{1i}$ and $\mathbf{b}_2$ in a state for deciding the bias vector $\mathbf{b}$. λ is called an association factor and to evaluate the bias vector eventually, we must find the association factor λ. There are many methodologies for finding the association factor λ, but we use an EM algorithm and a steepest-descent method that are evaluated by one iteration.

2.1.1. The case of an EM algorithm

We will describe the process of finding the association factor λ in the cepstral domain. The procedure of the EM-driven PMC with an association factor is as follows.

① Initialize the association factor. In this paper, the association factor is determined experimentally as 0.2.

② Perform initial PMC for input noisy speech. The PMC compensates both mean vectors and covariance matrices of clean speech Gaussian mixtures.

③ Prerecognition : The compensated model parameters are used for preliminary speech recognition.

④ Segmentation of a top 1 candidate : A top 1 recognition candidate found in step 3 is segmented by using a Viterbi decoding algorithm.

⑤ A calculation of λ using an EM algorithm: For a new association factor $\bar{\lambda}$, we use the state sequences obtained in step 4 and an EM algorithm.

Using the new $\bar{\lambda}$ given step 5 and the equation (2), we can adapt the clean speech model parameters to the noisy speech. The compensated mean vector of Gaussian mixture is,

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i^o + \Delta\boldsymbol{\mu}_i \quad (3)$$

Where the $\boldsymbol{\mu}_i^o$ is the clean speech mean vector of the i-th Gaussian mixture trained by HMM, and the $\Delta\boldsymbol{\mu}_i$ is a variation vector in the i-th Gaussian mixture due to the noisy environment of the recognizer. We can define the $\Delta\boldsymbol{\mu}_i$ as the bias vector $\mathbf{b}$, then from the equation (2), (3), the vector $\hat{\boldsymbol{\mu}}_i$ is

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i^o + (1-\lambda)\mathbf{b}_{1i} + \lambda\mathbf{b}_2 \quad (4)$$

Now, we can define the Q-function for finding a new association factor λ as follows.

$$\mathbf{Q}(\lambda,\bar{\lambda}) = \sum_{t-0}^{T-1}\sum_{k=0}^{M-1} p(k \mid \mathbf{y}_t, \lambda, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{b_{1k}}, \mathbf{b_2}) \cdot \log p(\mathbf{y_t}, k \mid \bar{\lambda}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{b_{1k}}, \mathbf{b_2}) \quad (5)$$

where T is total length of an input test word and M is a number of Gaussian mixtures in a state. We desire an association factor $\bar{\lambda}$ maximizing the equation (5). Using $\mathbf{Q}(\lambda,\bar{\lambda})$ in (5), the gradient becomes

$$\frac{\partial \mathbf{Q}(\lambda,\bar{\lambda})}{\partial\bar{\lambda}} = \sum_{t=0}^{T-1}\sum_{k=0}^{M-1} \gamma_t(k)[(\mathbf{b_2}-\mathbf{b_{1k}})\ \boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_k^o - \lambda(\mathbf{b_2}-\mathbf{b_{1k}}) - \mathbf{b_{1k}})] \quad (6)$$

where $\gamma_t(k)$ represents *a posteriori probability*, and is defined as follows.

$$\gamma_t(k) = \frac{w_k N_k(\mathbf{y_t} \mid M)}{\sum_{l=0}^{M-1} w_l N_l(\mathbf{y_t} \mid M)} \quad (7)$$

where $M = (\lambda, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{b_{1k}}, \mathbf{b_2})$ is a HMM's model and $\mathbf{y_t}$ is an observation vector in time t. $N_l(\mathbf{y_t} \mid M)$ is an output probability of the l-th Gaussian mixture, and w_l is the mixture weight. By equating the eq. (6) to zero, we can obtain the desired $\bar{\lambda}$.

2.1.2. A case of a steepest-descent algorithm

To find a new association factor $\bar{\lambda}$, a steepest-descent algorithm can be used instead of the EM algorithm. When a HMM is given, we desire to get a new association factor $\bar{\lambda}$ that maximizes the output probability of the observation vector. A probability P that an observation vector $\mathbf{y_t}$ is an outcome from a given HMM is

$$P = \prod_{t=0}^{T-1}\sum_{k=0}^{M-1} w_k N(\mathbf{y}_t \mid M = (\lambda, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{b_{1k}}, \mathbf{b_2})) \quad (8)$$

If we take logarithm in both sides of the equation (8), the gradient with respect to λ becomes

$$\frac{\partial \log P}{\partial\lambda} = \sum_{t=0}^{T-1}\sum_{k=0}^{M-1} \gamma_t(k) \cdot (\mathbf{b_2}-\mathbf{b_{1k}})^t\ \boldsymbol{\Sigma}_{\mathbf{k}}^{-1}(\mathbf{y_t} - \boldsymbol{\mu}_k) \quad (9)$$

where t is a vector transpose and the mean vector $\boldsymbol{\mu}_k$ is

$$\boldsymbol{\mu}_k = \tilde{\boldsymbol{\mu}}_k + \lambda(\mathbf{b_2}-\mathbf{b}_{1k}) + \mathbf{b}_{1k} \quad (10)$$

where $\tilde{\boldsymbol{\mu}}_k$ represents the original mean vector trained by clean speech.

Then, an estimated association factor $\bar{\lambda}$ is

$$\bar{\lambda} = \lambda - \beta \cdot \frac{\partial \log P}{\partial\lambda} \quad (11)$$

where λ is the previously estimated association factor , β is learning rate determining convergence speed. Although the adaptation may be repeated for convergence, in this paper one

iteration is taken to reduce the recognition time. And β is experimentally determined.

Once more differentiation of the equation (9) produces

$$\frac{\partial^2 \log P}{\partial \lambda^2} = -\sum_{t=0}^{T-1}\sum_{k=0}^{M-1} \gamma_t(k) \cdot (\mathbf{b_2} - \mathbf{b_{lk}})^t \ \mathbf{\Sigma_k^{-1}} (\mathbf{y_t} - \boldsymbol{\mu}_k) \qquad (12)$$

where covariance matrices are diagonal and positive definite. Because the equation (12) is always negative, the log-likelihood of the equation (8) has a maximum value.

2.2. The EM-driven VTS0 approach in the cepstral domain

In this section, we explain a model parameter compensation algorithm using 0th order VTS approximation with cepstral feature vectors. The clean speech is assumed to be corrupted by additive noise and spectral tilt. The environment model is as follows.

$$\mathbf{Y} = \left|\mathbf{H}(w)\right|\mathbf{X} + \mathbf{N} \qquad (13)$$

where $\mathbf{X}, \mathbf{Y}$ and $\mathbf{N}$ represents respectively a clean speech vector , a noisy speech vector, and a noise vector. The transfer function $\mathbf{H}(w)$ represents spectral tilt. Let us define the noise, $\mathbf{n}$ and spectral tilt vector, $\mathbf{h}$ as unknown random vectors in the cepstral domain. We assume that they are statistically independent of the clean speech vector $\mathbf{x}$. By taking logarithm and DCT (discrete cosine transform) to both sides of equation (13), we obtain

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{C} \cdot \log(1 + 10^{\mathbf{C}^{-1}(\mathbf{n}-\mathbf{h}-\mathbf{x})}) = \mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{n}, \mathbf{h}) \qquad (14)$$

where $\mathbf{x}, \mathbf{y}$ is respectively a clean speech vector and noisy speech vector in the cepstral domain and $\mathbf{n}, \mathbf{h}$ is respectively a noise vector and spectral tilt vector in the cepstral domain. $\mathbf{C}$ represents a DCT matrix. If we take a 0th order VTS approximation in equation (14), the mean and covariance of the noisy speech vector $\mathbf{y}$ becomes

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}} &= \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{f}(\boldsymbol{\mu}_{\mathbf{x}}, \bar{\mathbf{n}}_{\mathbf{0}}, \bar{\mathbf{h}}_{\mathbf{0}}) \\ \boldsymbol{\Sigma}_{\mathbf{y}} &\approx \boldsymbol{\Sigma}_x \end{aligned} \qquad (15)$$

Because the vectors $\bar{\mathbf{n}}$ and $\bar{\mathbf{h}}$ are expectations of unknown random vectors, we will estimate the vector $\bar{\mathbf{n}}$ and $\bar{\mathbf{h}}$ using an EM algorithm. First, the noise vector $\bar{\mathbf{n}}$ is estimated.

The Q-function used for estimating a new noise mean vector $\bar{\mathbf{n}}$ is

$$Q(\mathbf{n}, \bar{\mathbf{n}}) = \sum_{t=0}^{T-1}\sum_{k=0}^{M-1} p(k \mid \mathbf{y}_t, \mathbf{n}, M) \log p(\mathbf{y_t}, k \mid \bar{\mathbf{n}}, M) \qquad (16)$$

where $M = (\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \mathbf{h})$ and the vector $\bar{\mathbf{h}}$ is a known vector. Both sides of the equation (16) are differentiated by the vector $\bar{\mathbf{n}}$ and we can find the vector $\bar{\mathbf{n}}$ maximizing the Q-function as follows.

$$\bar{\mathbf{n}} = \frac{\sum_t \sum_k \boldsymbol{\gamma}_t(k) \Sigma_{\mathbf{x},k}^{-1} (\mathbf{y}_t - \mathbf{b}_k)}{\sum_t \sum_k \boldsymbol{\gamma}_t(k) \Sigma_{\mathbf{x},k}^{-1}} \qquad (17)$$

where $\mathbf{b}_k = \mathbf{C} \cdot \log(1 + 10^{\mathbf{C}^{-1}(\boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{h} - \mathbf{n})})$,

$$\boldsymbol{\gamma}_t(k) = \frac{w_k N_k(\mathbf{y}_t, \mathbf{n}, M)}{\sum_{l=0}^{M-1} w_l N_l(\mathbf{y}_t, \mathbf{n}, M)}$$

where $\mathbf{n}$, $\mathbf{h}$ is respectively initial value of the noise vector $\mathbf{n}$ and the spectral tilt vector $\mathbf{h}$. And $\boldsymbol{\mu}_{\mathbf{x},k}$ is the mean vector of the k-th Gaussian mixtures of the clean speech vector $\mathbf{x}$. Also, the expectation of the spectral tilt vector $\mathbf{h}$ can be estimated through the same procedures as the noise vector $\mathbf{n}$, but at this time, noise vector $\mathbf{n}$ is assumed to have a known value. The estimated spectral tilt vector $\bar{\mathbf{h}}$ is

$$\bar{\mathbf{h}} = \frac{\sum_t \sum_k \boldsymbol{\gamma}_t(k) \Sigma_{\mathbf{x},k}^{-1} (\mathbf{y}_t - \mathbf{a}_k)}{\sum_t \sum_k \boldsymbol{\gamma}_t(k) \Sigma_{\mathbf{x},k}^{-1}} \qquad (18)$$

where $\mathbf{a}_k = \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{C} \cdot \log(1 + 10^{\mathbf{C}^{-1}(-(\boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{h} - \mathbf{n}))})$

Similarly as above, we can also estimate the environment parameters using the 1st order VTS approximation.

3. EXPERIMENTAL RESULTS

3.1. The database and feature parameters

In our experiments, we evaluate the performance of the proposed algorithms when there is a mismatch between the training and testing environments. For isolated word recognition experiments, we use a left-to-right continuous density-HMM(CHMM)[3] with 3 states. And the database consist of 75 phoneme-balanced words[4]. The speech samples are recorded in a silent office environment, and converted by A/D converter with 16KHz sampling rate and 16bit quantization[5]. The train database is composed of speech material uttered by 15 speakers, and the test database is constructed by 5 speakers excluded from the training. The noise sources used in recognition experiments are additive white Gaussian noise and car noise recorded at a 90-120[km/h] speed in an express highway. To get corrupted speech we added noise to clean speech at various signal-to-noise ratios (SNR). In this experiments, we use 13 standard Mel-frequency cepstral coefficients including a normalized frame energy and their derivatives as the feature parameters. The HMM recognizer is based on 32 phone like unit (PLU).

3.2. Baseline experiments

The experimental result of the baseline recognizer without model parameter compensation in noisy speech recognition is as follows.

	clean	30dB	20dB	10dB	0dB
AWG	93.9	87.2	54.9	24.3	4.7
CAR	93.9	94.4	89.6	63.2	26.4

Table 1. The result of the baseline recognizer without model parameter compensation

From the above result, we can see that the recognition rate deteriorates suddenly below 20dB and 10dB. Furthermore, the recognition rate in 0dB AWG approaches zero. The recognition result with matched training and testing environments are shown in table 2. We can show that the results are superior to any of them in table 1. Besides, the recognition rate in 0dB AWG case is improved by about ten times. This experimental result may be approximately the upper limit of recognition rate. However, this requires retraining of the recognizer to fit in new environments, which may not be possible in real situation. Therefore, we need recognition methods which is robust to noisy environments without retraining. Next, we discuss experimental results of the proposed model parameter compensation algorithms.

	30dB	20dB	10dB	0dB
AWG	94.4	91.5	85.3	56.3
CAR	93.9	93.9	90.1	89.6

Table 2. The result of the same environment in training and testing

3.3. Experimental results for model parameter compensation algorithms

In this experiment, we use the Jack's knife method for more reliable performance comparison. MPMC-EM and MPMC-STP are the modified PMC methods using respectively EM and steepest-descent algorithms for an association factor. EM-VTS0 is the 0th order mean-VTS using an EM algorithm.

	Clean	30dB	20dB	10dB	0dB
PMC	91.4	90.5	81.9	65.0	29.6
MPMC-EM	91.7	90.9	87.0	70.3	31.7
MPMC-STP	91.6	91.1	86.7	70.8	31.5
VTS-0	92.6	91.0	87.0	68.8	27.6

Table 3. The result of AWG noise

	Clean	30dB	20dB	10dB	0dB
PMC	91.4	90.5	89.8	85.9	72.5
MPMC-EM	91.7	92.1	90.9	87.8	78.0
MPMC-STP	91.6	91.7	90.7	86.8	73.2
VTS-0	92.6	92.2	91.3	88.9	77.6

Table 4. The result of CAR noise

From the above result, we can see that the MPMC and VTS-0 show better recognition results than the conventional PMC in all SNR. Also, the algorithms show better recognition results for the case of CAR noise than the AWG noise. Especially, The proposed model compensation algorithms show better performances than conventional PMC in low SNR.

4. SUMMARY

In this paper, we proposed new model parameter compensation algorithms for noise-robust speech recognition in the cepstral domain. They have shown good recognition performance, and are simply realized.

The recognition rate of the MPMC without covariance compensation is higher than the conventional PMC. This confirms that the nonuniform model parameter compensation is better suited in noisy speech recognition.

The recognition rate of mean-VTS0 is higher than PMC due to good modeling of environments by reflecting additive noise and spectral tilt. More appropriate modeling of noisy environments may lead to better performances.

5. REFERENCE

[1] M. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise". *in Proceedings of ICASSP*, pp. 233- 236, 1992.

[2] M. Gales and S. Young, "Cepstral parameter compensation for HMM recognition". *Speech Communication*, no. 3, pp. 233-236, 1993.

[3] L. R. Rabiner and B. H. Juang, "An Introduction to hidden Markov Models". *IEEE ASSP Magazine*, no. 1, pp. 4-16, Jan. 1986.

[4] C. K. Un, D. Y. Kim, and Y. H. Chang, *A Study on Noise-Robust Techniques for Speech Recognition.* Final Report, KAIST, Dec. 1996.

[5] I. J. Choi, O. W. Kwon, J. L. Park, D. Y. Kim, H. Y. Chung, C. K. Un, "Korean speech database for auto-interpretation", *Proceedings of speech communication and signal processing workshop*, pp. 287-290, 1994. (Edited by Korean)

[6] P. Moreno, *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon Univ., April 1996.

[7] M. Gales, *Model-based techniques for noise robust speech recognition*. PhD thesis, Gonville and Caius College, Sep. 1995.