A New Distance Measure for a Variable-Sized Acoustic Model Based on MDL Technique

Hoon-Young Cho and Sanghun Kim

Embedding a large vocabulary speech recognition system in mobile devices requires a reduced acoustic model obtained by eliminating redundant model parameters. In conventional optimization methods based on the minimum description length (MDL) criterion, a binary Gaussian tree is built at each state of a hidden Markov model by iteratively finding and merging similar mixture components. An optimal subset of the tree nodes is then selected to generate a downsized acoustic model. To obtain a better binary Gaussian tree by improving the process of finding the most similar Gaussian components, this paper proposes a new distance measure that exploits the difference in likelihood values for cases before and after two components are combined. The mixture weight of Gaussian components is also introduced in the component merging step. Experimental results show that the proposed method outperforms MDL-based optimization using either a Kullback-Leibler (KL) divergence or weighted KL divergence measure. The proposed method could also reduce the acoustic model size by 50% with less than a 1.5% increase in error rate compared to a baseline system.

Keywords: Acoustic modeling, optimization, minimum description length, parameter reduction.

Manuscript received Mar. 5 2010; revised July 6, 2010; accepted July 21, 2010.

This work was supported by the Ministry of Knowledge Economy, Korea (2010-S-019-03, Development of Portable Korean-English Automatic Speech Translation Technology).

Hoon-Young Cho (phone: +82 42 860 6591, email: hycho@etri.re.kr) and Sanghun Kim (email: ksh@etri.re.kr) are with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.10.1510.0062

I. Introduction

Most contemporary speech recognizers are based on hidden Markov models (HMM). In continuous speech recognition tasks with large vocabularies, several tens of thousands of words and their pronunciations form a dictionary. HMM-based acoustic phone models are trained to obtain the statistical distributions of acoustic instances of each phone. In the decoding step, an input speech signal is represented as a sequence of feature vectors, and a series of words are searched from a network that is composed of a whole acoustic model set, a pronunciation lexicon, and a language model.

In many large-vocabulary continuous speech recognition systems, tied-state context-dependent triphone models have been used, where the number of unique HMM states ranges from 2,000 to 6,000, each of which is a mixture of about 8 to 64 Gaussian components. Since the likelihood score of each HMM state should be calculated at every frame in the decoding step, it is reported that 30% to 70% of the total recognition time is spent by the likelihood estimation [1]. Therefore, there have been many studies on reducing the number of Gaussian components in HMM states with a minimal loss of recognition accuracy. As more ASR systems are being developed for mobile devices where memory size and computational power are limited, an efficient reduction of acoustic model size is becoming more important [2].

Among several previous well-known studies related to the acoustic model reduction problem, the semi-continuous HMM technique shares a codebook of Gaussian mixture components across all models [1]. A variant of this approach is merging the Gaussian components in such a way that likelihood loss is minimized, and only the covariance terms are shared. It is reported that when the covariances are tied, not only is the

parameter size reduced, but the covariance also becomes more robust [3]. A greedy clustering method iteratively finds and merges a pair of Gaussian components with the smallest value of the cost function until a target number of Gaussians is reached [4]. Similarly, the Gaussian components of each state are clustered into a binary tree, and a subset of the components is chosen from the Gaussian tree on the basis of the minimum description length (MDL) criterion [5]-[8]. This technique could reduce the model size by 50% to 75% with only a slight degradation in accuracy. The MDL criterion is closely related to the Bayesian information criterion and Akaike information criterion. The differences and similarities between them are discussed in [8]. Another direction of optimization might be sub-vector clustering techniques that tie the model parameters even at a granularity finer than a Gaussian component, reducing the amount of quantization errors [9], [10].

In the framework of MDL-based optimization, one of the major issues is distance measures for finding a pair of the closest components when building a binary Gaussian tree in each state. Several distance measures such as Kullback-Leibler (KL) divergence, Bhattacharyya distance, and weighted KL (wKL) divergence have been compared. The wKL divergence gave the best performance among these, as the other two measures neglect the weight term of a Gaussian mixture component [5]-[8], [11].

Once a binary tree is constructed and fixed, the next step only traverses the tree nodes and selects their optimal subset. Therefore, the topology of the resultant binary Gaussian tree is very important. The topology is determined by the distance measure used, along with the component merging method.

Since the main purpose of choosing a better distance measure is to find a pair of components that gives the minimum difference in likelihood of the state before and after merging, we propose another distance measure that directly exploits the likelihood change and thus improves the topology of binary Gaussian trees.

The rest of this paper is organized as follows. Section II describes the overall optimization procedure of an acoustic model based on the MDL technique. Section III discusses previous similarity measures for binary mixture component tree building and the proposed method. Section IV reviews the MDL criterion for pruning the binary tree in order to optimally reduce the model parameters. In section V, we compare the performances of the proposed method with the previous approaches. Finally, section VI concludes this paper.

II. Overview of Model Parameter Reduction

As mentioned earlier, many acoustic model optimization methods first build a best model using as many parameters as possible. After that, some of the redundant parameters are either removed or merged according to a given criterion [6], [12].

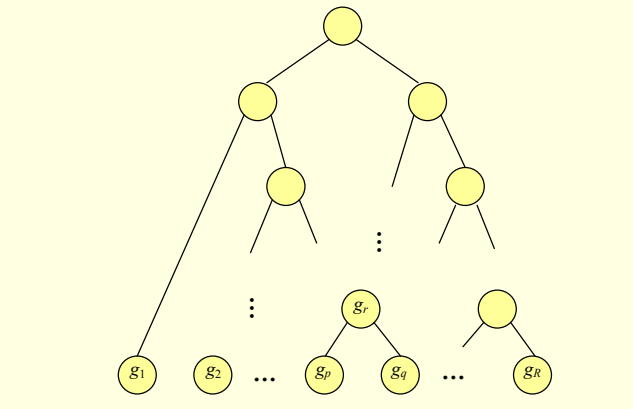


Fig. 1. Binary Gaussian mixture tree built at an HMM state.

Let us assume that there are R Gaussian mixture components in an arbitrary HMM state as shown in Fig. 1. Among all possible pairs of mixture components, the closest pair should be found and then merged. At this stage, we need to have an appropriate similarity measure and component merging method. Let the two closest components be g_p and g_q, which are merged into a parent node g_r. The closest pair is searched again from the remaining R–2 components and the newly generated parent node. This process continues until we have a single root node.

Once a binary tree is constructed, the next step is pruning the tree according to the MDL criterion. Beginning from the root node, the process calculates the description length change of splitting a node into its children nodes. If the difference between description lengths before and after the splitting is positive, it means that the total description length is becoming larger. Therefore, the splitting is stopped at the node. Otherwise, the splitting repeats for each of the children nodes until no nodes remain to process.

A binary tree is constructed for each HMM state, and optimization on the number of Gaussian components in each state is performed in such a way that the total number of Gaussian mixture components included in an entire acoustic model meets a predetermined value.

The performance of this optimization technique depends on the quality of the binary tree constructed in each HMM state and an effective pruning technique. The former is very important, because once a tree is built, the pruning process has no choice but to follow the edges of the tree and cut off some redundant branches.

Distance measures and an estimation of the model parameters used for merged components are the two main factors influencing the quality of a binary tree. Therefore, in the following section, we discuss the previous distance measures and the component merging method used in building a binary

tree. Then, we propose a new distance measure and component merging method to improve binary tree quality.

III. Distance Measures and Component Merging

1. Conventional Distance Measures

Let s be an HMM state that is composed of M Gaussian mixture components. The likelihood score of an observation feature vector $\mathbf{x}$ is calculated as

$$\Pr(\mathbf{x} \mid s) = \sum_{m=1}^{M} G_m(\mathbf{x}) = \sum_{m=1}^{M} w_m \cdot g_m(\mathbf{x}) = \sum_{m=1}^{M} w_m \cdot N(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m), \quad (1)$$

where w_m, $\boldsymbol{\mu}_m$, and $\boldsymbol{\sigma}_m$ are the mixture weight, mean vector, and covariance matrix of the m-th Gaussian mixture component, respectively.

If we denote the two arbitrary Gaussian components by $G_p(\mathbf{x})$ and $G_q(\mathbf{x})$, the KL divergence between the two distributions is defined as

$$\begin{aligned} & d_{\mathrm{KL}}(G_p(x), G_q(x)) \\ &= \int g_p(x) \log \frac{g_p(x)}{g_q(x)} dx + \int g_q(x) \log \frac{g_q(x)}{g_p(x)} dx \\ &= \frac{1}{2} \sum_{d=1}^{D} \left[\frac{\sigma_p^{\,2}(d) + (\mu_p(d) - \mu_q(d))^2}{\sigma_q^{\,2}(d)} \right. \\ &\qquad \left. + \frac{\sigma_q^{\,2}(d) + (\mu_q(d) - \mu_p(d))^2}{\sigma_p^{\,2}(d)} \right], \end{aligned} \quad (2)$$

where D indicates the dimension of the feature vector.

Recently, the weight term of a Gaussian mixture component was added to (2) [5]. This wKL divergence is defined as

$$\begin{aligned} & d_{\mathrm{wKL}}(G_p(x), G_q(x)) \\ &= \int w_p g_p(x) \log \frac{w_p g_p(x)}{w_q g_q(x)} dx + \int w_q g_q(x) \log \frac{w_q g_q(x)}{w_p g_p(x)} dx \\ &= \frac{1}{2} \sum_{d=1}^{D} (w_q - w_p) \log \frac{\sigma_p^{\,2}(d)}{\sigma_q^{\,2}(d)} \\ &\quad + \frac{1}{2} \sum_{d=1}^{D} \left[\frac{w_p \{\sigma_p^{\,2}(d) + (\mu_p(d) - \mu_q(d))^2\}}{\sigma_q^{\,2}(d)} \right. \\ &\qquad \left. + \frac{w_q \{\sigma_q^{\,2}(d) + (\mu_q(d) - \mu_p(d))^2\}}{\sigma_p^{\,2}(d)} \right] \\ &\quad + D\{(w_p - w_q) \log w_p + (w_q - w_p) \log w_q - 0.5(w_p + w_q)\}. \end{aligned} \quad (3)$$

As is mentioned in [5], it is necessary to find a pair of mixture components that gives the minimum difference in likelihood before and after the two components are merged. It is obvious that the likelihood calculation will be more accurate when the KL divergence considers the mixture weights, as in (3).

2. Conventional Component Merging Method

After the two closest Gaussian components are found, the pair should be merged into a single Gaussian component to form a parent node. The mean and covariance of the new component [6] are obtained by

$$\begin{aligned} \mu_k(i) &= \frac{1}{M_k} \sum_{m=1}^{M_k} E(x_m^{(k)}(i)) = \frac{1}{M_k} \sum_{m=1}^{M_k} \mu_m^{(k)}(i), \\ \sigma_k^2(i) &= \frac{1}{M_k} \sum_{m=1}^{M_k} E((x_m^{(k)}(i) - \mu_k(i))^2) \\ &= \frac{1}{M_k} \left[\sum_{m=1}^{M_k} \sigma_m^{(k)^2}(i) + \sum_{m=1}^{M_k} \mu_m^{(k)^2}(i) - M_k \mu_k^{\,2}(i) \right], \end{aligned} \quad (4)$$

where M_k=2 because only two components are merged in the binary tree case. If the tree type is not restricted to a binary tree, the value for M_k is the number of children nodes for a given node.

3. Proposed Distance Measure and Merging Method

Let us recall that the purpose of a distance measure is to find a pair of Gaussian components that shows the least change in likelihood scores after the two components are merged. We propose another similarity measure that directly exploits the change in likelihood score before and after the component merging. We call it a delta-likelihood (DL) distance measure.

Assuming that $\mathbf{X}_p = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $\gamma_p(\mathbf{x})$ are a set of feature vectors aligned to a Gaussian component g_p and the occupancy count of a feature vector at the component, respectively, the log likelihood can be obtained as follows [3].

$$\begin{aligned} LL(\mathbf{X}_p \mid g_p) &= \sum_{i=1}^{N} \gamma_p(\mathbf{x}_i) \log \Pr(\mathbf{x}_i \mid g_p) \\ &= -0.5 \cdot \gamma_p \cdot (D \log 2\pi + \log |\sigma_p| + D), \end{aligned} \quad (5)$$

where D is the dimension of the feature vectors, $\gamma_p = \sum_{i=1}^{N} \gamma_p(\mathbf{x}_i)$, and σ_p is the covariance of the Gaussian component.

Assuming also that two Gaussian components g_p and g_q are merged into g_r, the difference of the log likelihood before and after the merging can be expressed as

$$\begin{aligned} \Delta &= LL(\mathbf{X}_p \mid g_p) + LL(\mathbf{X}_q \mid g_q) - LL(\mathbf{X}_r \mid g_r) \\ &= -0.5(\gamma_p \log |\sigma_p| + \gamma_q \log |\sigma_q| - (\gamma_p + \gamma_q) \log |\sigma_r|). \end{aligned} \quad (6)$$

We propose using the above quantity as the cost of merging the two Gaussian components. If the cost is small, it means the two

components are close enough to be combined. Since the occupancy values, γ_p and γ_q, are often not available, the proposed DL measure uses the Gaussian mixture weights, w_p and w_p, instead, because they have similar meaning. The proposed DL distance measure is defined as

$$d_{\mathrm{DL}}(G_p(x), G_q(x)) = (w_p + w_q)\log|\sigma_r| - w_p \log|\sigma_p| - w_q \log|\sigma_q|. \quad (7)$$

The proposed distance measure always has a zero or positive value since the likelihood score is larger when a given data is represented with twice the parameters.

Finally, (8) through (10) are the proposed component merging method for any number of Gaussian components. Unlike the previous merging method, the proposed merging technique considers weight terms of the Gaussian mixture.

$$\alpha_m^{(k)} = w_m^{(k)} \Big/ \sum_{m=1}^{M_k} w_m^{(k)}, \quad (8)$$

$$\mu_k(i) = \sum_{m=1}^{M_k} \alpha_m^{(k)} E(x_m^{(k)}(i)) = \sum_{m=1}^{M_k} \alpha_m^{(k)} \mu_m^{(k)}(i), \quad (9)$$

$$\begin{aligned}\sigma_k^2(i) &= \sum_{m=1}^{M_k} \alpha_m^{(k)} E\left[(x_m^{(k)}(i) - \mu_k(i))^2\right] \\ &= \sum_{m=1}^{M_k} \left(\alpha_m^{(k)} E\left[(x_m^{(k)2}(i)\right]\right) - 2\mu_k(i)\mu_k(i) + \sum_{m}^{M_k} \alpha_m^{(k)} E\left[\mu_k{}^2(i)\right] \\ &= \sum_{m=1}^{M_k} \left[\alpha_m^{(k)} \sigma_m^{(k)2}(i) + \alpha_m^{(k)} \mu_m^{(k)2}(i)\right] - \mu_k{}^2(i). \quad (10)\end{aligned}$$

Comparing (8) through (10) with (4), we notice that the previous merging method is a special case of applying $\alpha_m^{(k)} = 1/M_k = const.$ into (9) and (10).

IV. MDL Criterion for Tree Pruning

Once a binary tree is constructed, the description length for each subset of all the tree nodes is calculated, and the node set which has an MDL is selected. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a series of data, and $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be a set of estimated model parameters to represent data $\mathbf{x}$. The MDL criterion function [7], [13] is defined as

$$MDL(\mathbf{X}) = \min_{\lambda,k} \left\{ -\log P_\lambda(\mathbf{X}) + \alpha \cdot \frac{k}{2} \log N + C \right\}. \quad (11)$$

Because the probability is higher as the modeling power is increased for a given data set, the first term on the right side of the equation will decrease as the complexity of the model increases. In the second term, k is the number of model parameters. In HMM-based speech recognition, k is the total number of Gaussian mean vectors and covariance matrices. This term increases as the model has more parameters and works as a penalty for increasing the model complexity. The α value controls the degree of penalty. The last term, C, is a constant. Figure 2 shows three examples of the MDL criterion function. We can see that the optimum points of the graphs are determined by the α value.

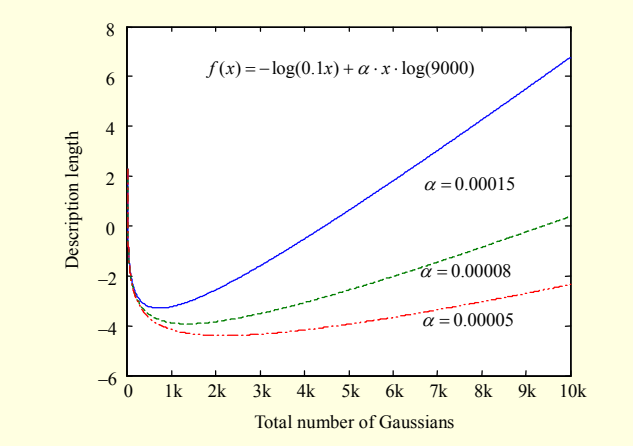


Fig. 2. MDL criterion functions for three penalty values: optimal number of total Gaussian components is different.

V. Experiments

To evaluate the proposed and previous optimization methods, we performed speech recognition experiments on the travel domain. HMM models were trained using version 3.4.1 of the HMM toolkit (HTK). Multiple speech databases related to the travel domain including the Wall Street Journal and TIMIT database were used for the acoustic model training. The total size of the training data was about 330 hours. We extracted 39 dimensional mel-frequency cepstral coefficient features. Tied-state context-dependent triphone HMMs were generated through the model training process.

The baseline system consisted of 17,159 physical triphone models, where the number of unique states was 4,554. Since we used 16 Gaussian components in each state, the total number of Gaussian components was 72,864. We used 776,000 sentences to build a 3-gram language model. The HDecode utility from the HTK was used for the recognition experiments. Lastly, the test data were made up of 200 sentences from a travel-domain speech database collected in a laboratory environment.

Table 1 presents the speech recognition performances for four different methods used when building a binary tree at each state: the baseline and three MDL-based optimization methods using KL divergence, wKL divergence, and the proposed DL distance measures. In this experiment, no extra iterations of

Table 1. Word accuracies of baseline system, KL, wKL, and DL distance measures in MDL-based optimization. Performances without retraining optimized models.

Average no. of mixtures/state	Baseline	KL	wKL	DL (proposed)
16	81.51	–	–	–
12	79.45	80.59	81.02	81.13
8	69.47	78.58	78.52	78.63
4	62.26	69.74	72.78	74.35

Table 2. Word accuracies of baseline system, KL, wKL, and DL distance measures in MDL-based optimization. Performances after three iterations of model retraining.

Average no. of mixtures/state	Baseline	KL	wKL	DL (proposed)
16	81.51	–	–	–
12	79.45	81.13	81.29	81.62
8	69.47	79.18	79.72	79.12
4	62.26	72.45	74.67	75.22

model training were done after the optimized models were generated. All the comparisons in this paper were fulfilled using the same weighted component merging method. Since the number of Gaussian components is variable among HMM states, the table shows the word accuracies when the averaged numbers of mixture components are 16, 12, 8, and 4.

As we reported, the wKL divergence showed a better optimization performance than the KL divergence. This means that the quality of binary trees that are built at each state is better with the wKL measure than with the KL measure. The proposed DL distance measure showed an improved performance over the wKL because it better optimizes the difference in likelihood values while binary Gaussian trees are generated.

Table 2 shows the word accuracies when three iterations of maximum likelihood retraining were performed after the new models had been generated. Model retraining can be considered when the training database is available at the model optimization step, though in some cases, the training data are no more accessible. The results show that after three iterations, the performances are not degraded much when the model parameters are reduced to 75% of the original model. The proposed method could reduce the acoustic model size by 50% with less than a 2.4% increase in error rate compared to the baseline system. Compared with Table 1, the performances were improved after the retraining.

Figure 3 shows the word accuracies of the baseline

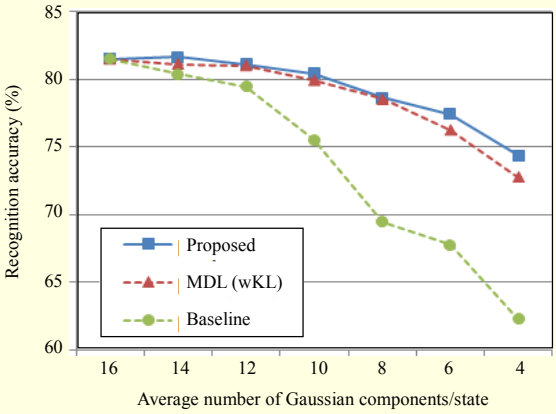


Fig. 3. Word accuracies of proposed method, conventional MDL-based method using wKL distance, and baseline system over various numbers of mixture components.

Table 3. Word accuracies of wKL and DL distance measures in MDL-based optimization using a WFST-based decoder. The accuracy is 88.73% and averaged RTF is 0.86 for a baseline system.

Average no. of mixtures/state	Total no. of mixtures	wKL	DL (proposed)	Average RTF
16	79,424	–	–	0.86
12	59,568	88.18	88.34	0.74
8	39,713	86.98	87.26	0.59
4	19,856	84.78	85.69	0.34

recognition system, the previous MDL-based method using the wKL divergence, and the proposed method as the average number of Gaussian components per state is decreased. Both of the MDL-based optimization methods improve the word accuracy of the baseline system by more than 10% when the model size is downsized to a quarter of the original system. The DL method consistently showed a higher performance than the previous wKL divergence in the framework of MDL-based optimization.

Finally, we evaluated the proposed method using a larger test set. The test set consisted of 1,446 English sentences. The average sentence length was 9.13 words. In this experiment, we used a one-pass decoder based on a weighted finite-state transducer instead of the HTK in order to examine the dependency of the proposed method on different decoders. We used the same training data for the baseline acoustic model as in the above experiments. The real-time factors (RTF) of the test utterances were also measured to check the effectiveness of the proposed method.

The results are summarized in Table 3. The proposed method

showed better accuracy than the wKL measure. Compared with the HDecode of the HTK, the performance using a WFST-based one-pass decoder was less degraded. Since the averaged RTFs of the two methods were almost the same, we show the RTFs of only the proposed method. It is clear that the overall recognition speed improves with the reduction of Gaussian mixture components.

VI. Conclusion

In this paper, we proposed a delta-likelihood distance measure and a weighted component merging method in the framework of minimum-description-length (MDL)-based model parameter optimization. Experimental results showed that the proposed method could reduce the acoustic model size by 50% with less than a 1.5% increase in error rate in comparison to the baseline system. Furthermore, it consistently showed higher performances than the previous MDL-based method using the weighted KL divergence distance measure.

References

[1] J. Cai et al., "Efficient Likelihood Evaluation and Dynamic Gaussian Selection for HMM-based Speech Recognition," *Comput. Speech Language*, vol. 23, 2009, pp. 147-164.

[2] I.L. Hetherington, "PocketSUMMIT: Small-Footprint Continuous Speech Recognition," *Proc. INTERSPEECH*, 2007, pp. 1465-1468.

[3] M.Y. Hwang and X. Huang, "Dynamically Configurable Acoustic Models for Speech Recognition," *Proc. ICASSP*, 1998, pp. 669-672.

[4] P.L. Dognin et al., "Refactoring Acoustic Models using Variational Density Approximation," *Proc. ICASSP*, 2009, pp. 4473-4476.

[5] A. Ogawa and S. Takahashi, "Weighted Distance Measure for Efficient Reduction of Gaussian Mixture Components in HMM-Based Acoustic Model," *Proc. ICASSP*, 2008, pp. 4173-4176.

[6] K. Shinoda and K. Iso, "Efficient Reduction of Gaussian Components Using MDL Criterion for HMM-Based Speech Recognition," *Proc. ICASSP*, 2002, pp. 869-872.

[7] K. Shinoda and T. Watanabe, "MDL-Based Context-Dependent Subword Modeling for Speech Recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, 2000, pp. 99-102.

[8] K. Shinoda, "Robust Acoustic Modeling for Speech Recognition," *IEICE Technical Report*, vol. 104, no. 541, 2004, pp. 7-12.

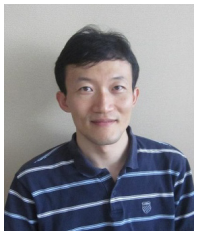
[9] E. Bocchieri, "Vector Quantization for the Efficient Computation of Continuous Density Likelihoods," *Proc. ICASSP*, 1993, pp. 692-695.

[10] G.J. Jung, H.Y. Cho, and Y.H. Oh, "Data-Driven Subvector Clustering Using the Cross-Entropy Method," *Proc. ICASSP*, 2007, pp. 977-980.

[11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* 2nd Ed., Wiley Interscience, 2000.

[12] G.F.G. Yared, F. Violaro, and L.C. Sousa, "Gaussian Elimination Algorithm for HMM Complexity Reduction in Continuous Speech Recognition System," *Proc. INTERSPEECH*, 2005, pp. 377-380.

[13] J. Rissanen, "Universal Coding, Information, Prediction, and Estimation," *IEEE Trans. IT,* vol. 30, 1984, pp. 629-636.



Hoon-Young Cho received the BS and MS in computer science from KAIST, Daejeon, Korea, in 1995 and 1998, respectively, and the PhD in electrical engineering and computer science from KAIST, Daejeon, Korea, in 2003. He was a visiting scholar at UC, San Diego, in 2004. He was a senior researcher at the Mobile Multimedia Research Center of LG Electronics from November 2004 to January 2006. Since 2006, he has been with the Department of Speech and Language Information Research at ETRI, Daejeon, Korea. Currently, he is a senior researcher in the Automatic Speech Translation Research Team. His interests include robust speech recognition, machine learning, and multimedia signal processing.

Sanghun Kim received the BS in electrical engineering from Yonsei University, Seoul, Korea, in 1990, and the MS degree in electrical and electronic engineering from KAIST, Daejeon, Korea, in 1992. He received his PhD from the Department of Electrical, Electronic, Information, and Communication Engineering at the University of Tokyo, Japan, in 2003. Since 1992, he has been with the Research Department of Spoken Language Processing Section of ETRI, Daejeon, Korea. Currently, he is a principal researcher in the Automatic Speech Translation Research Team. His interests include speech synthesis, speech recognition, and speech signal processing.