Optimization of Gaussian Mixture in CDHMM Training for Improved Speech Recognition

Seo-gu Lee[*] · Sung-gil Kim[*] · Sun-mee Kang[**] · Han-seok Ko[**]

ABSTRACT

This paper proposes an improved training procedure in speech recognition based on the continuous density of the Hidden Markov Model (CDHMM). Of the three parameters (initial state distribution probability, state transition probability, output probability density function (p.d.f.) of state) governing the CDHMM model, we focus on the third parameter and propose an efficient algorithm that determines the p.d.f. of each state. It is known that the resulting CDHMM model converges to a local maximum point of parameter estimation via the iterative Expectation Maximization procedure. Specifically, we propose two independent algorithms that can be embedded in the segmental K-means training procedure by replacing relevant key steps; the adaptation of the number of mixture Gaussian p.d.f. and the initialization using the CDHMM parameters previously estimated. The proposed adaptation algorithm searches for the optimal number of mixture Gaussian humps to ensure that the p.d.f. is consistently re-estimated, enabling the model to converge toward the global maximum point. By applying an appropriate threshold value, which measures the amount of collective changes of weighted variances, the optimized number of mixture Gaussian branch is determined. The initialization algorithm essentially exploits the CDHMM parameters previously estimated and uses them as the basis for the current initial segmentation subroutine. It captures the trend of previous training history whereas the uniform segmentation decimates it. The recognition performance of the proposed adaptation procedures along with the suggested initialization is verified to be always better than that of existing training procedure using fixed number of mixture Gaussian p.d.f.

Keywords : CDHMM, mixture gaussian

* School of Electrical Engineering, KOREA UNIVERSITY

** Department of Computer Science, SEOKYEONG UNIVERSITY

1. INTRODUCTION

Hidden Markov Models (HMMs) have been demonstrated as one of the most powerful statistical tools available for automatic speech recognition [1]. The characteristic parameters of the employed Markov process model is usually estimated by the maximum likelihood (ML) method, which presumes that the amount of training data is large enough to provide accurate estimates [2][3]. HMMs can be based on either discrete output probability distributions (e.g., discrete HMM) or continuous output probability density functions (e.g., semi-continuous density HMM (SCDHMM) or CDHMM).

In the discrete HMM, the discrete probability distributions are sufficiently powerful to model any random events with a reasonable number of parameters. The major problem of the discrete output probability is that the vector quantization partitions the acoustic feature into separate regions through some distortion measure. This raises a problem in that the partition operations may destroy the original signal structure. Many studies indicate that the recognition accuracy for discrete HMM is indeed lower than that of the CDHMM [4][5][6]. To overcome this limitation, either the SCDHMM or the CDHMM can be chosen as the candidate for investigation. In the HMMs of continuous output probability density functions, the parameter estimation of speech model is usually based on the ML methods on the assumption that observed signal is generated by a mixture Gaussian process [7]. Continuous density HMM captures the essence of feature vectors accurately and has the interpolating ability in spectral modeling while providing flexibility.

In the HMM-based parameter estimations for speech recognition, modeling the variation of spectral characteristics in speech signal is a crucial problem. There are many known factors contributing to increase in the variation. Context phonemes surrounding a center phoneme, speaker identity, speaking rate, signal power, accent and background noise, are some of the examples among them. In HMM, output distributions model the various spectral characteristics in each state.

In this paper, an effective estimation method of each states output p.d.f. is proposed. The output p.d.f. is an essential parameter in CDHMM which can significantly affect the recognition performance. Our aim is to improve the acoustic modeling in HMM in two aspects. First, in the modeling process, more tangible efforts should be made to take into account the different spectral characteristic of each HMM state. Second, in the initialization process, more effective estimates are helpful in discrete symbol cases are essential in continuous distribution cases.

As solution to the above two problems, we propose an efficient adaptation of the number of mixture Gaussian p.d.f., and an effective initialization using the CDHMM parameters previously estimated, respectively. Section 3 describes these two algorithms in detail. Both approaches collectively provide an effective ML estimation of CDHMM by incorporating the spectral characteristics of each state. They are embedded into the segmental K-means training procedure, which is widely used for parameter estimation of CDHMM, and the recognition performance is evaluated.

The remainder of this paper is organized as follows. In Section 2, we present the conventional re-estimation algorithm of continuous density HMM. In Section 3, we discuss the estimation method using variable number of mixture Gaussian p.d.f. and suggest an effective initialization algorithm. We then apply the algorithm to Korean isolated word database for automated voice dialing system (VDS). The experiment arrangements and the recognition results are discussed in Section 4. Finally, we summarize our findings in Section 5.

2. RE-ESTIMATION ALGORITHM OF CONTINUOUS DENSITY HMM

The most difficult problem of HMMs is to determine an effective method that adjusts the model parameters $\lambda=(A,B,\pi)$ to satisfy a certain optimization criterion. There is no known way to analytically solve for the model parameter set that maximizes the probability of the observation sequence in a close form [9]. The two well-known iterative techniques used to estimate the model parameters in CDHMM are the Baum-Welch algorithm and segmental K-means algorithm respectively. In this section, we discuss the two algorithms to formulate the problem to be pursued.

2.1. Baum-Welch reestimation method

Baum and his colleagues have determined that either (1) the initial model λ defines a ciritical point of the likelihood function, in which case $\bar{\lambda}=\lambda$; or (2) model $\bar{\lambda}$ is more likely than model λ in the sense that $P(O\mid\bar{\lambda})>P(O\mid\lambda)$; such that a new model $\bar{\lambda}$, from which the observation sequence is more likely, is produced. Based on this premise, if we use $\bar{\lambda}$ in place of λ and repeat the reestimation calculation, we can improve the probability of O being observed from the model until some training point is reached. The final result of the reestimation procedure is the ML estimate of the HMM.

The reestimation formulas can be derived directly by maximizing Baum's auxiliary function,

$$Q(\lambda',\lambda)=\sum_{q}P(O,\ q\mid\lambda')\log P(O,\ q\mid\lambda) \tag{1}$$

over λ , because

$$Q(\lambda',\lambda) \geq Q(\lambda',\lambda') \Rightarrow P(O \mid \lambda) \geq P(O \mid \lambda') \tag{2}$$

We can maximize the fumction $Q(\lambda',\lambda)$ over λ to improve λ' in the sense of increasing the likelihood $P(O \mid \lambda)$. Eventually the likelihood function converges to a local maximum point if we iterate the procedure [9].

2.2. Segmental K-means reestimation method

An alternative way to train the HMM parameters by the ML criterion is the segmental K-means algorithm [10]. The segmental K-means algorithm is a procedure for estimating the HMM parameters by embedding the K-means method into a Markov chain. The segmentation information can be obtained from the Viterbi decoding procedure [11]. Contrary to Baum's algorithm, the segmental K-means algorithm provides an estimate, which locally maximizes the joint likelihood of the observation sequence and the most likely state sequence. Instead of likelihood function $P(O \mid \lambda)$, $\max_s P(O, s \mid \lambda)$ is used as the optimization objective. The motivation for using $\max_s P(O, s \mid \lambda)$ as the optimization criterion is reasonable [6][12]. It was found that the likelihood values associated with the parameter sets estimated by the two algorithms are very close and, futhermore, the estimated parameter sets themselves are similar [13]. It is known that both algorithms converge to local maximum points. We employ the segmental K-means algorithm because it can be implemented simply by segmenting and clustering speech signals, and can avoid the numerical difficulties associated with Baum's algorithm.

In the segmental K-means algorithm, model parameters are estimated to maximize joint probability for observing the sequence O along with the most likely state sequence s.

$$\bar{\lambda} = \arg\max_{\lambda}\left[\ \max_s P(O, s \mid \lambda)\right] \tag{3}$$

To accomplish this objective, the EM algorithm is used in two steps. In the first step, the optimal state sequence that maximizes the probability of the observation sequence of model $\bar{\lambda}$ is obtained.

$$\bar{s} = \arg\max_s P(O, s \mid \lambda) \tag{4}$$

Then, based on the state sequence $\bar{s}$, a new model $\overline{\lambda_{n+1}}$ is estimated by

$$\overline{\lambda_{n+1}} = \arg\max {}_{\lambda} P(\boldsymbol{O}, \ \bar{\boldsymbol{s}} \mid \lambda) \tag{5}$$

3. ESTIMATION USING VARIABLE NUMBER OF MIXTURE GAUSSIAN P.D.F. AND SUGGESTED INITIALIZATION

In this section, we formulate the use of variable number of mixture Gaussian p.d.f. for the estimation of CDHMM parameters.

When we consider the sequence of feature vectors, we expect that certain portions of a speech utterance are often more useful in classifying the utterance than other portions. In other words, certain portions of a speech utterance have more distinguishing features than other portions. Therefore, the performance of the recognizer is expected to be improved by imposing appropriate weights to the HMM states proportional to their utterance characterizing values [8]. There have been many attempts to show that the score of a speech utterance through a valid state sequence is a weighted sum of HMM log state-likelihood through the state sequence [14]. The state-weighting approaches, on the other hand, involve many complex procedures and intractable mathematical equations.

This paper proposes a simpler approach by focusing on the variable number of mixture Gaussian p.d.f. of each state and its optimization, instead of state-weights, to reflect upon the characteristics of each state and the variety of speech utterances. Usually the number of mixture components for each state is constant throughout all the states. Since such a scheme ignores the characteristics of each state, it results in a coarse modeling of speech signal.. To model speech signal as accurately as possible, we propose to direct a distinctive number of mixture components for each state.

There can be many criteria that establish the optimal number of mixture Gaussian p.d.f. in each state. We use one component of the HMM parameters. Among the HMM parameters $\lambda=(A, B, \pi)$, we choose the use $B_i=(\pi_{im,} \Sigma_{im}, c_{im})$ as the criterion in state s_i. In particular, the parameter Σ_{im} can be considered as a performance measure of the estimation efficiency. It is reasonable to propose that the p.d.f. of large value Σ must have more mixture branches than that of small value Σ.

Two forms of the Σ matrices can be considered; namely, the diagonal matrices (with assumed zero correlation between components of the representation), and full covariance matrices. The advantage of the diagonal covariance matrix is that the computation of $b_j(o)$ reduces to a simple sum of products of Gaussians, whereas for a full covariance matrix, the computation $b_j(o)$ requires a matrix multiplication. On the other hand, the disadvantage of the diagonal covariance matrix representation is that, in general, for correlated vector components, a larger value of M (the number of mixtures) is needed to establish an adequate model than for a full covariance matrix representation [4]. This paper chooses to employ the diagonal matrix. The determinant of m_{th} diagonal matrix in state j is

$$D_{jm} = \prod_{k=1}^{p} \Sigma_{jm}(k) \tag{6}$$

where p is the dimension of speech feature vector.

We can consider that the determinant D_{jm} is a measure of dispersion about Σ. Because of the multiple Gaussian humps in a mixture Gaussian density, the dispersion criterion on state i can be considered as a collection of weighted sum of the determinants:

$$CRITERION = \sum_{m=1}^{M} c_{im} D_{im} \tag{7}$$

This way, we also can regulate the number of mixture using the change of *CRITERION*.

In a typical training procedure, there is one outstanding problem we all encounter. In theory, the reestimation algorithm can guarantee that the HMM parameters correspond only to the local maximum of likelihood function. Finding the optimal initial estimate of the HMM parameters makes it possible that the local maximum is the global maximum of the likelihood function. Basically there is no simple or straightforward answer [9].

The initialization algorithm that uses uniform segmentation is the simplest method. To improve the recognition performance, we can consider first each speech feature vector as a possible target for improvement. For example, we can calculate the distance between contiguous frames and use this information as the segmentation criterion for initialization. But such an initialization approach is complex and independent of training parameters. Instead, we propose an initialization procedure that utilizes the CDHMM parameters previously estimated. Subsection 3.2 shows this procedure in detail. It is embedded in the segmental K-means training algorithm for measuring its effectiveness. It is a convenient method that makes it unnecessary to consider each speech feature vector.

3.1. Finding the optimal number of mixture Gaussian p.d.f.

The proposed algorithm follows the steps outlined below. First, the number of mixtures in each state *(m)* is selected. We apply the segmental K-means training algorithm and compute the *CRITERION*. This procedure is repeated after increase of *m* until the all convergence conditions of each state are satisfied.

- Stepl : Initialization

  Linearly segment all training vectors into phoneme or word units and HMM states. By clustering, the parameters a_{ij} and $(\mu_{im}, \Sigma_{im}, c_{im})$ are initialized.

- Step2 : Segmentation

  The CDHMM parameters estimated in Step 1 or Step 3 are used to (re)segment each training utterance into phoneme units and the HMM states via Viterbi decoding. The transition probabilities are obtained using the segmentation information from Viterbi decoding. That is

$$a_{ij} = \frac{n_{ij}(s^*)}{\sum_{k\geq 1} n_{ik}(s^*)} \tag{8}$$

  where s^* is the most likely state sequence and $n_{ij}(s^*)$ denotes the count of transition from state s_i to state s_j for s^*..

- Step 3 : Clustering and Estimation

  All the observation vectors corresponding to a partial state of each phoneme model are partitioned into M clusters using the standard vector quantization (VQ) design method, and the parameters $(\mu_{im}, \Sigma_{im}, c_{im})$ are estimated for each cluster m $(M \geq m \geq 1)$ in state s_i as

$$\mu_{im} = \frac{1}{L_{im}} \sum_{O_t \in V_{im}} O_t \tag{9}$$

$$\Sigma_{im} = \frac{1}{L_{im}} \sum_{O_t \in V_{im}} (O_t - \mu_{im})(O_t - \mu_{im})^T \tag{10}$$

$$c_{im} = \frac{L_{im}}{\sum_{k=1}^{M} L_{jk}} \tag{11}$$

  where V_{im} denotes a set of vectors that have been partitioned to the m_{th} mixture of state s_j and L_{im} denotes the number of members in V_{im}. Figure 1 shows this step.

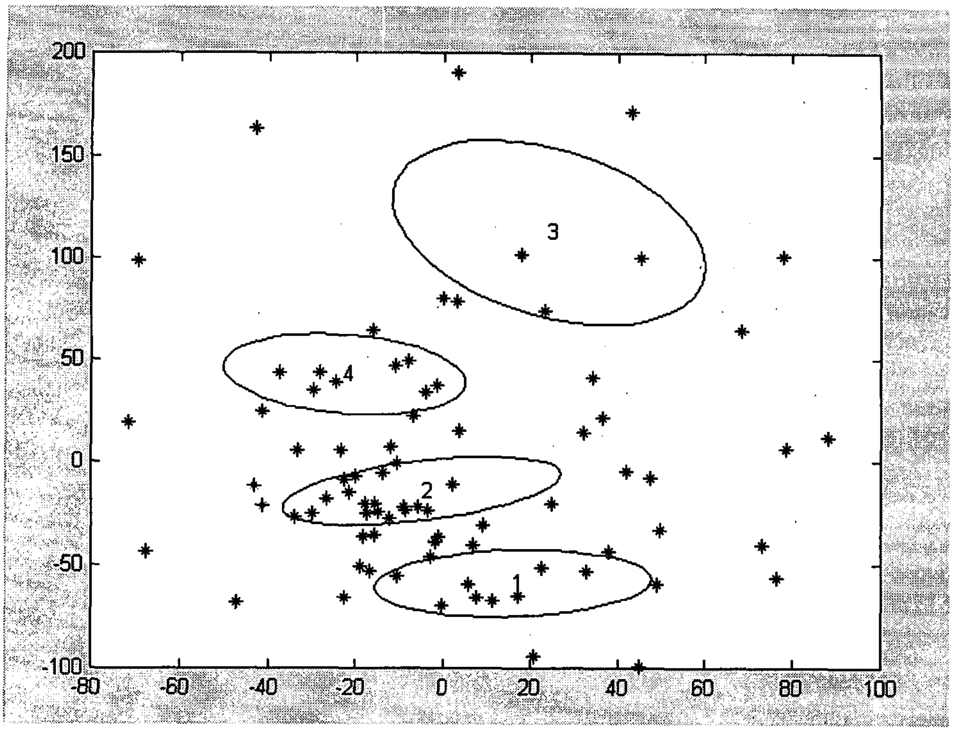

Figure 1. An example of clustering and estimation using feature vectors 2-dimension

- Step 4 : Decision

  Compute $D_{jm}=\prod_{k=1}^{p}\Sigma_{jm}(k)$ at each state and mixture

  And then, compute $CRITERION=\sum_{m=1}^{M}c_{im}D_{im}$

  if *CRITERION* converges, *m* is fixed
  or else *m* increases

- Step 5 : Repetition

  Steps *1- 4* are repeated until all of the
  convergence conditions are satisfied.

3.2. Initialization algorithm using the CDHMM parameters previously estimated

The initialization procedure we propose is based on the parameter-dependent segmentation instead of uniform segmentation. This procedure can be implemented by modifying *Step 1* and *Step 4* of the proposed algorithm in subsection 3.1. To avoid the error from poorly estimated parameters arising from small *m*, *m* is set to a large number initially. In this procedure, the change of *m* is in a decreasing order as opposed to the increasing order as represented in subsection 3.1.

- Step 1 : Initialization

  Segment all training vectors into the HMM states using the CDHMM parameters previously estimated.

  By clustering, the parameters a_{ij} and $(\mu_{im}, \Sigma_{im}, c_{im})$ are initialized.

- Step 4 : Decision

  Compute $D_{jm}= \prod_{k=1}^{p} \Sigma_{jm}(k)$ at each state and mixture.

  And then compute $CRITERION= \sum_{m=1}^{M} c_{im}D_{im}$

  If *CRITERION* converges, *m* is fixed

  else *m* decreases and the current parameters are stored.

4. EXPERIMENTAL RESULTS

To study the effect of the proposed method, various experiments are conducted. An isolated word recognition system that recognizes 50 Korean words is implemented for experiments. The baseline system (Figure 2) consists of Feature Extraction, Training and Recognition using the Hidden Markov Model (HMM).

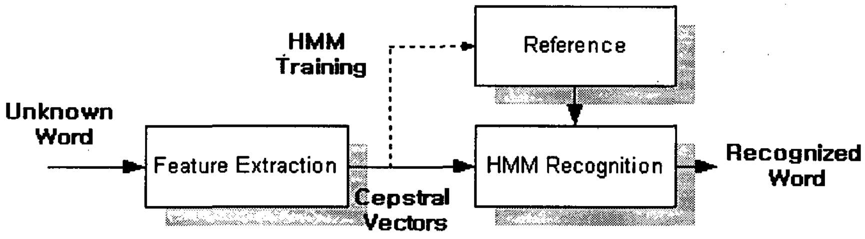

Figure 3. Baseline Speech Recognition System

In the front end, a feature extraction algorithm is applied to speech sampled at 8 kHz and produces coefficient in spectral domain, i.e. cepstral vectors. In addition to static information provided by the cepstral vectors, many systems typically use dynamic information represented by differences of the cepstral vectors. These feature vectors are :

- 12 static cepstral coefficients,

- 12 1^{st} order differential cepstral coefficients,
- 12 2^{nd} order differential cepstral coefficients,
- 1^{st} order and 2^{nd} order differences of the log power,

Differences of cepstral vectors are given by :

$$d_t = \frac{\sum_{k=1}^{K} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \tag{12}$$

where c_t is cepstral vectors and d_t is its difference at frame t [15].

Only 12 static cepstral coefficients are used in the experiment for convenience.

The baseline system is an isolated word recognizer. It is designed to recognize 50 Korean words. Each word is modeled by a CDHMM with a 5-state left-right model (Figure 3) [4].

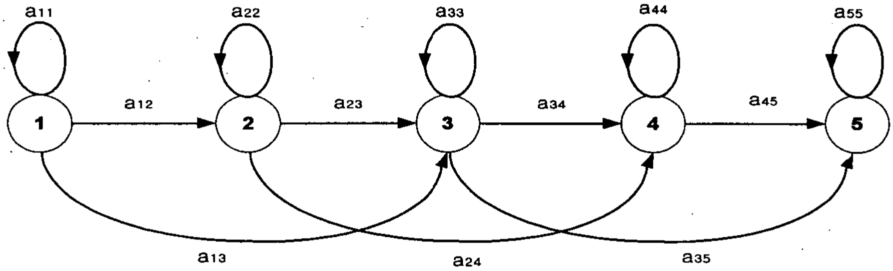

Figure 3. Left-right HMM.

Our training and recognition tests are performed using the database for the Voice Dialing System (VDS). Collected via telephone line, the database is sampled at 8 kHz and band-limited to the frequency range from 300 to 3,400 Hz. The database consists of 13 male speakers in their twenties. The speakers have pronounced each word five times. Among the database, 2,495 words from 10 speakers are used for training and 749 words from 3 speakers are used for recognition testing.

4.1. Comparison of variable branch with fixed branch

The proposed method is compared with the algorithm that has fixed number of mixture. In the performance test, we've observed the recognition rate of a test word using topsl, tops2 and tops3. Topsl represents the matching score with the most likely candidate while tops2 represents the matching score up to the 2nd most likely candidate. And tops3 represents the matching score up to the 3rd most likely candidate.

First, Figures 4, 5 ,6 show the convergence of *CRITERION* as iteration number increases. We can regulate the number of mixture using the change of *CRITERION;*

$$CRITERION = \sum_{m=1}^{M} c_{im} D_{im} \tag{13}$$

The *CRITERION* is reduced as the iteration number increases. This way, we can limit the mixture number when the *CRITERION* falls below the threshold. Table I shows a comparison of the proposed training method applying the variable branch with the method of fixed branch when training iteration was performed five times. The threshold value is $0.006*10^{10}$. The average number of variable mixture density is found to be 5.41. It is noted that the performance with variable number of Gaussian mixture components is better than that of using the fixed number of Gaussian distributions.

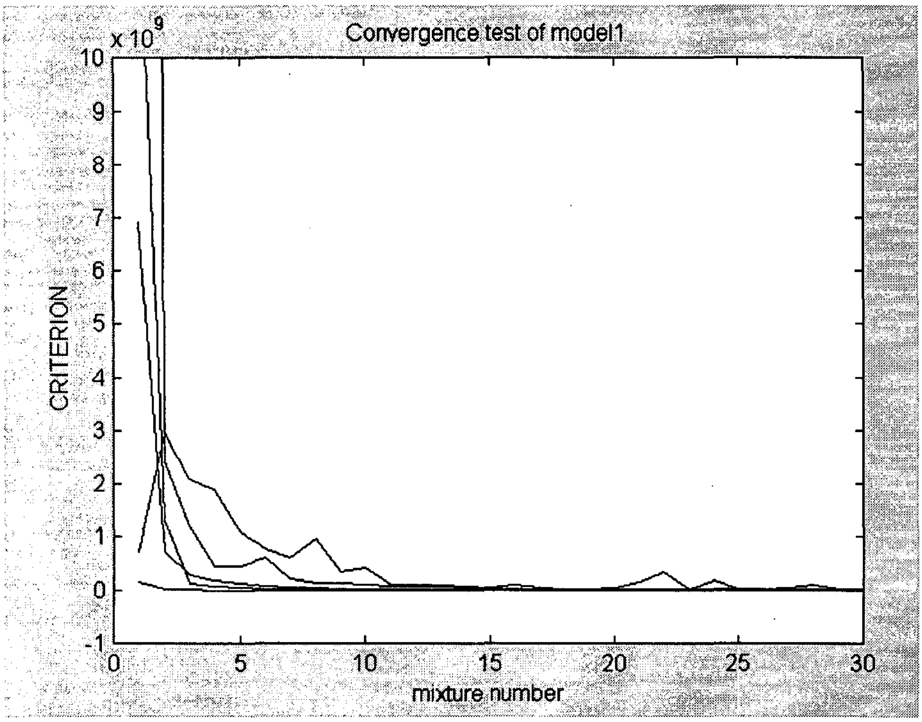

Figure 4. Example of convergence test (model 1)

A comparison in terms of noisy speech shows a negative effect. Table II shows the proposed training method applied variable branch with the method of fixed branch when SNR is 10 dB.

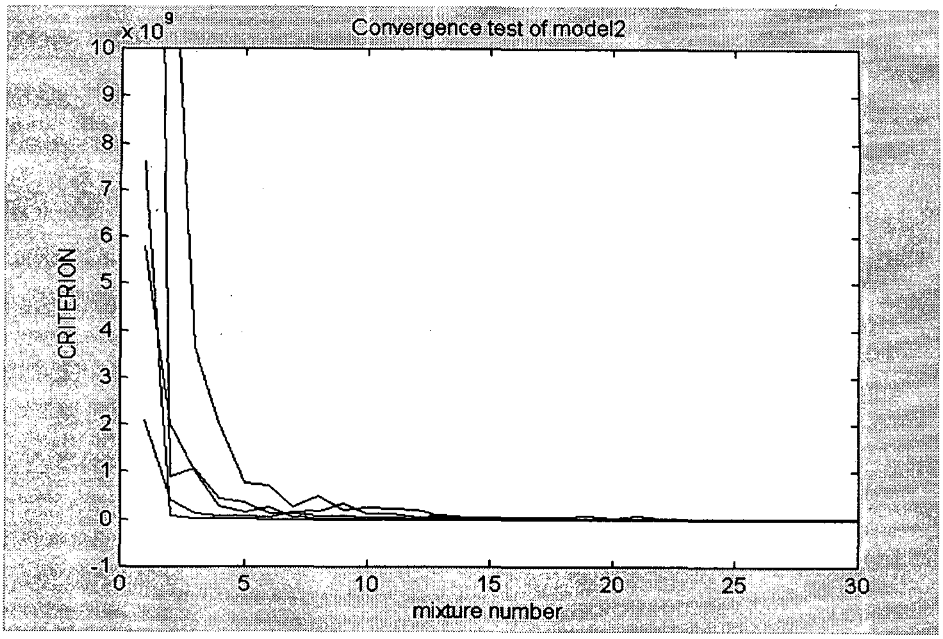

Figure 5. Example of convergence test (model 2)

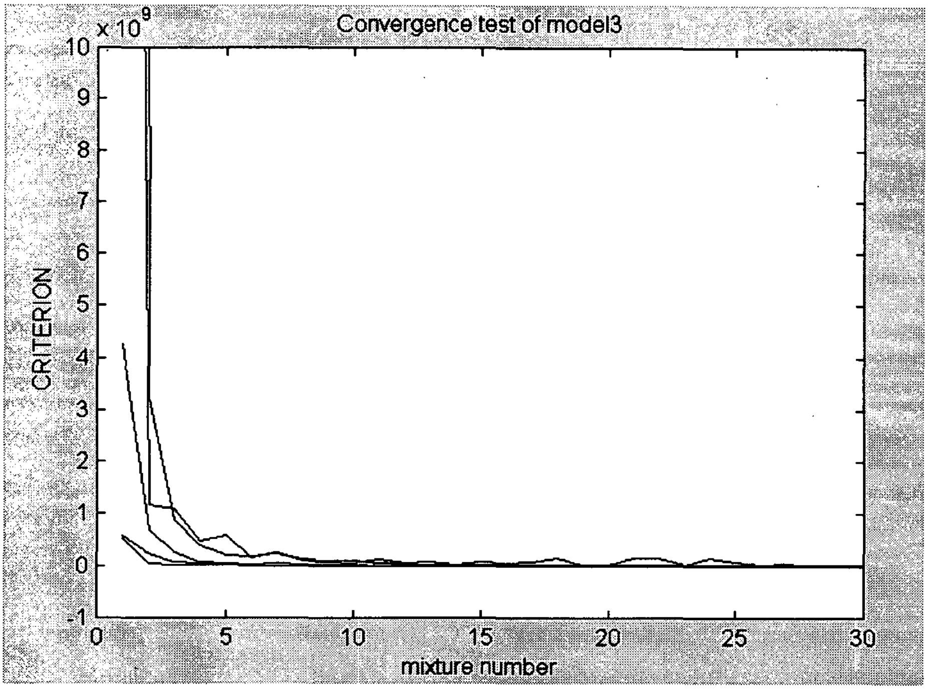

Figure 6. Example of convergence test (model 3)

Table I. Recognition rates with respect to training method. Iter=5

Training method		Variable branch (5.41)	Fixed branch (6)
Recognition	top1	95.59	95.19
Rate	top2	98.80	98.26
%	top3	99.47	99.47

Table II. Recognition rates with respect to training method. SNR=10dB, Iter=5

Training method		Variable branch (5.41)	Fixed branch (6)
Recognition	top1	74.37	76.50
Rate	top2	85.58	86.87
%	top3	89.85	90.65

4.2. Results of initialization based on the CDHMM parameters previously estimated

Table III shows recognition rates based on initialization using the CDHMM parameters estimated in previous training procedure versus initialization by uniform segmentation. The results show that the initialization procedure using proposed algorithm is better than the conventional method.

Table III. Recognition rates of initialization using parameters previously estimated. Iter=5.

Training method		Initialization using proposed algorithm	Unifrom initialization
Recognition	top1	95.99	95.59
Rate	top2	98.93	98.80
%	top3	99.60	99.47

5. CONCLUSION

In this paper, we propose an improved training procedure in speech recognition based on continuous density Hidden Markov Model (CDHMM). The algorithm can be embedded in the segmental k-means training algorithm by replacing the appropriate steps.

The local maximum of parameter estimation is efficiently attained by the two proposed procedures, the adaptation of the number of mixture Gaussian p.d.f. and the initialization using splitting of mixture Gaussian branch. By applying this technique using a threshold value of the variance of state, the optimized number of mixture Gaussian branch is estimated.

The performance of the proposed method was tested on a speaker-independent isolated word system that recognizes 50 Korean words using speech database for voice dialing service. The comparison of the proposed method with other methods shows that the performance of the proposed method is superior to the conventional methods.

REFERENCES

[1] LEE, K.. 1989. *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic, Boston, MA.

[2] GAUVAIN, J. L. and LEE, C. H. 1994. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE Trans. Speech, Audio processing* , vol. 2, 291-298.

[3] BAUM, L. E., PETRIE, T., SOULES, G., and WEISS, N. 1970. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." *Ann. Math. Stat.*, 164-171.

[4] RABINER, L. R., JUANG, B. H., LEVINSON, S. E., and SONDHI, M. M.. 1985. "Recognition of isolated digits using Hidden Markov Models with continuous mixture densities." *AT&T Bell Lab. Tech. J.*, 1211-1234.

[5] HUANG, X. D. and JACK, M. A. 1989. "Semi-continuous Hidden Markov Models for speech signals," *Computer Speech and Language*, 234-251.

[6] PARK, Y. K. 1995. *A Study on the Use of Statistical Information in Speech Recognition based on HMM*, Ph.D. Thesis, Korea Advanced Institute of Science and Technology.

[7] HUANG, X. D., ARIKI, Y., and JACK, M. A. 1990. *Hidden Markov Models for Speech Recognition*, Edinburgh University Press.

[8] SU, K. Y. and LEE, C. H. 1994. Speech recognition using weighted HMM and subspace projection algorithm, *IEEE Trans. Speech, Audio processing*, vol. 2, 69-79.

[9] RABINER, L. and JUANG, B. H. 1993. Fundamentals of Speech Recognition, *Prentice-Hall International, Inc.*, 1993.

[10] RABINER, L. R., WILPON, J. G., and JUANG, B. H. 1986. "A segmental k-means training procedure for connected word recognition." AT&T Bell Lab. *Tech. J.*

[11] FORNEY, G. D. 1973. "The viterbi algorithm," *Proceedings of The IEEE*, **61**(3), 263-277.

[12] JUANG, B. H. and RABINER, L. R. 1990. "The segmental k-means algorithm for estimating parameters of Hidden Markov Models." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**(9), 1639-1641.

[13] MERHAV, N. and EPHRAIM. Y. 1991. "Maximum likelihood Hidden Markov Modeling using a dominant sequence of states." *IEEE Transactions on Signal Processing*, 39, 2111-2114.

[14] KWON, O. W. 1996. *On Improving Acoustic Modeling in Speech Recognition based on Continuous Density HMM*, Ph.D. Thesis, Korea Advanced Institute of Science and Technology.

[15] KIM, J. 1998. "Wavelet Transform based Feature Extraction for Speech Recognition." *The Journal of the Acoustical Society of Korea*, Vol. 2, 31-37.

접수일자 : '99. 2. 15.
게재결정 : '99. 3. 11.

▲ Seo-gu Lee
LG Electronics, Display Division
e-mail: seogulee@lge.co.kr

▲ Sung-gil Kim
School of Electrical Engineering
Korea University
Tel: (02) 927-6115
e-mail: sgkim@ispl.korea.ac.kr

▲ Sun-mee Kang
Department of Computer Science, Seokyeong University
Research interests include speech and character recognition.
e-mail: smkang@bukak.seokyeong.ac.kr

▲ Hanseok Ko
School of Electrical Engineering, Korea University
Research interests include speech, acoustics, and image signal processing.
Tel: (02) 927-6115
e-mail: hsko@ispl.korea.ac.kr