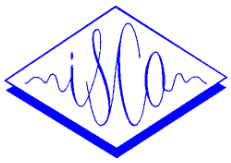

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

AN OPTIMAL DISCRIMINATIVE TRAINING METHOD FOR CONTINUOUS MIXTURE DENSITY HMMS

Shinobu Mizuta and *Kunio Nakajima*

Information Systems and Electronics Development Laboratory
Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura, 247 Japan

Abstract

In this paper, we describe a training method for continuous mixture density HMM parameters, called optimal discriminative training. Conventional maximum likelihood estimation method for HMM training has a problem that the recognition performance is not considered in the training procedure. To solve the problem, a corrective training method has been already proposed, but this method is applied to discrete HMMs, so the trained HMMs cannot avoid undesirable effects of VQ distortion. In this paper, the optimal discriminative training method (ODT) is described, appling the the basic concept of the corrective training to continuous mixture density HMMs, better recognition performance can be obtained as avoiding the VQ distortion. From word recognition experiments, we discuss the way to optimize each parameters of this training method, and by using the optimum value of parameters, we show the effectiveness of this method.

1. INTRODUCTION

Recently, for automatic speech recognition, a hidden Markov model(HMM) is widely used, representing a time sequence of acoustical features with a statistical way.

HMMs can be classified into two groups by expressive form: discrete HMMs and continuous density HMMs. Discrete HMMs cannot avoid the influence of VQ distortion, but the continuous density HMMs are free from the distortion, and higher recognition performance can be obtained especially when that is continuous mixture density HMMs[1]. Particularly, the latter is effectable for speaker-independent recognition which treats varied acoustic features.

Generally, training methods for HMM parameters are based on the maximum likelihood estimation (MLE) method, but the MLE method has a problem that the discriminative accuracy between the confusable categories is not considered in the training. To solve the problem for discrete HMMs, the corrective training method [2][3][4] has been proposed, and the improvement of recognition performance has been shown by some experiments.

In this paper, we describe a training method of HMM, called the optimal discriminative training (ODT) method [5]. The ODT method is similar to the corrective training method, but is altered applying to continuous mixture density HMMs, so as to exclude the VQ distortion.

We discuss the ability of the ODT method through the speaker-independent word recognition experiments. At first, we investigate the strategy to optimize each parameters of the ODT method. Then, we show the improvement of recognition performance by the ODT method.

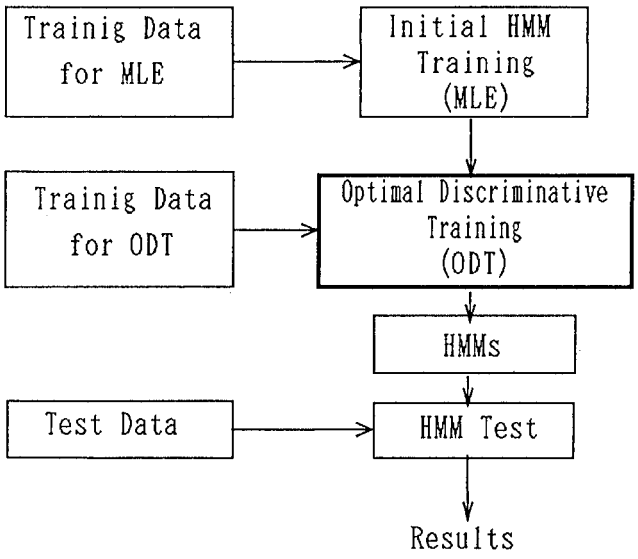

Fig.1 A block diagram of recognition using HMM based on ODT.

1. OPTIMAL DISCRIMINATIVE TRAINING

Figure 1 shows a block diagram of the recognition system using continuous mixture density HMMs based on ODT. In the ODT method, each sample of the training data set is recognized using initial HMMs. These HMMs are trained by MLE method. If the sample is misrecognized or nearly misrecognized the two HMMs are picked up: one belongs to the correct category, and the other belongs to the incorrect but most probable category. After that, the parameters of these two HMM are re-trained, so that the correct category is more probable and the incorrect category is less probable.

This HMM training method is realized by the following procedure.

1. Train the initial HMM set S with the MLE method.
2. For each sample in the training data, calculate the probabilities for the HMMs, using the Viterbi algorithm.
3. For a training sample L(c) of category c, pick up the HMM H(c) of correct category and H(n) of most probable incorrect category, and compute the difference D(c,n) between the probabilities P(c), P(n) for these two HMMs.

$$D(c,n) = P(c) - P(n) \tag{1}$$

4. For the i th frame of L(c), pick up the corresponding state s(c,i) of H(c) and s(n,i) of H(n), using the Viterbi pathes computed at step 2.
5. For the i th frame vector V(i) of L(c), pick up the most probable Gaussian density m(c,i) and m(n,i) from the output probabilities of S(c,i) and S(n,i) represented by mixturized Gaussian densities.
6. Move the mean vectors μ(c,i), μ(n,i) of m(c,i), m(n,i) by the following way:

$$\mu(c,i) = \mu(c,i) + \gamma(V(i) - \mu(c,i)) \tag{2}$$

$$\mu(n,i) = \mu(n,i) - \gamma(V(i) - \mu(n,i)) \tag{3}$$

Where,

$$\gamma = \begin{cases} 0 & (D(c,n) > \delta). \\ \beta & (D(c,n) < 0). \\ \beta(1 - D(c,n)/\delta) & (0 \le D(c,n) \le \delta). \end{cases} \tag{4}$$

$$(\beta > 0, \delta > 0) \tag{5}$$

7. Continue with step 2, until enough iterations have been run.

The way of moving the mean vectors on step 6 is similar to the way in the LVQ2[6].

3. EXPERIMENTAL EVALUATION

In this section, we discuss the performance of the ODT method in some experiments of speaker-independent word recognition.

At first, we investigate the effects of the ODT parameters on the recognition performance. It is desirable that the optimum value of each parameters can be obtained not to be based on recognition experiments, so we discuss here focusing on the optimizing method of the parameters without recognition. After that, we study the effectiveness of the ODT method.

3.1 Speech Representation

Speech signals are sampled at 10kHz, and pre-enphasized with a filter of $1 - 0.95z^{-1}$. Then, a Hamming window with a width of 25.6msec is applied every 10msec, and 15 mel-scale LPC cepstral coefficients and mean energy is computed. The acoustical features are comprised of these cepstral coefficients, regression coefficients of the cepstral coefficients, and regression coefficient of the mean energy[7].

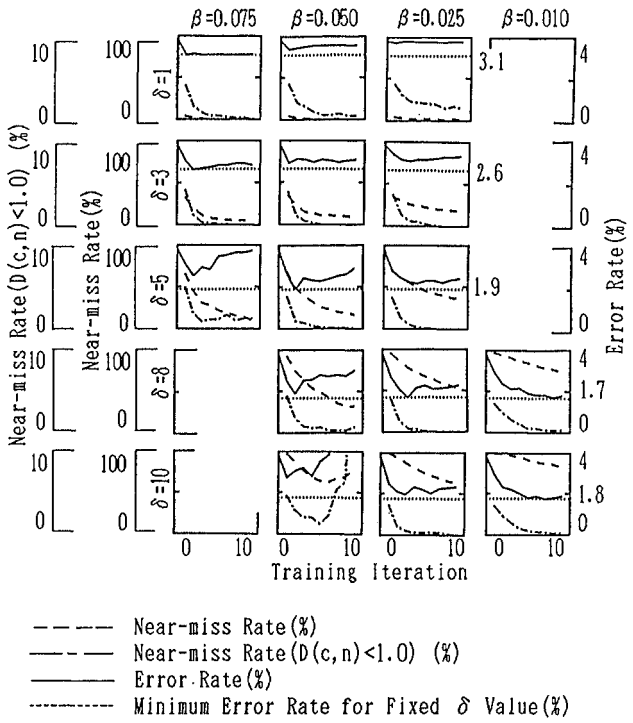

Fig.2 Near-miss rates, near-miss rates on condition $D(c,n) < 1.0$, and recognition error rates vs iteration for various near-miss criteria δ and learning rates β.

Word HMMs for the recognition experiments are comprised of sub-phonetic unit HMMs. As the sub-phonetic unit HMMs, the duration controllable HMMs [8] are used, and the output probabilities of the HMMs are represented by 4 mixture diagonal Gaussian densities.

As the speech database, Japanese common speech data corpus of JEIDA (JS-WRD-87)is applied. For experiments, data sets of 100 Japanese city names uttered once by 75 male speakers are used. The data set DST1 of 10 speakers and DST2 of 50 speakers are used as training data, and DSR1 of 25 speakers is used as test data. DST2 include DST1, and DSR1 consists of different speakers from DST2.

In the ODT iteration and the recognition, we selected only 10 confusable words for each sample to reduce the computation time.

3.2 Effects of Training Parameters

Near-miss criterion Figure 2 shows the relation between the recognition performance and the number of iteration for various near-miss criteria δ and learning rates β. Where, a ODT near-miss rate is a counting rate of near-miss samples ($D(c,n) < \delta$) and all training samples. In these experiments, error rates for training data (near-miss rates on condition $D(c,n) < 0$) take very few value, so we use near-miss rate on condition $D(c,n) < 1.0$, to indicate the recognition performance for the training data, instead of the error rate.

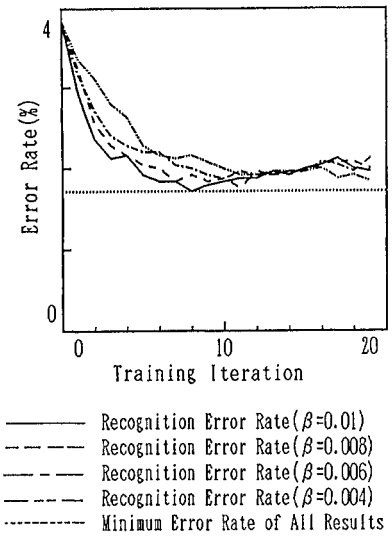

Fig.3 Recognition error rates vs iteration for various learning rates β (δ=8.0).

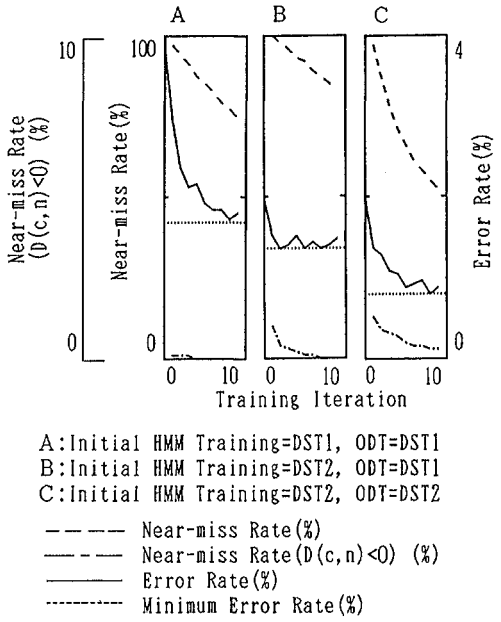

Fig.4 Near-miss rates, near-miss rates on condition $D(c,n) < 0$, and recognition error rates vs iteration for various training data sets (δ=8.0, β=0.01).

Most of near-miss rates and error rates for training data decrease with the ODT iteration. But, recognition error rate tend to increase after decreasing. It seems that the over-training of HMMs to the training data causes these phenomena.

The near-miss rate increases monotonously as δ increases. At $\delta \geq 8$, the near-miss rate of the first ODT iteration is almost 100%, and the HMMs can obtain the best recognition performance. This result demonstrates that we can determine the optimum value of δ by using near-miss rate. Therefore, the scheme determining optimal δ can realize without recognition experiments.

Learning rate When δ is fixed, the minimum recognition error rates in the ODT iterations tend to increase after decreasing as the value of β decreases. This phenomenon indicates that, even if a suitable β is selected for training data sets, excess or insufficient learning occurs for too large or too small β being applied to a recognition data, so β must be optimized in compliance with δ especially in case of smaller δ. As Figure 3 shows, when the value of δ is large enough, above mentioned increasing of minimum error rates doesn't occur. So, we can determine the value of β an appropriate small value as the optimum value.

Number of speakers The performance of the speaker-independent speech recognition greatly depends on the number of speakers for training. We discuss the number of speakers, for initial HMM training by MLE, and for ODT iteration.

Figure 4 shows the relation between the recognition performance and the number of iteration for two sets of HMM training data, DST1(10 male speakers) and DST2 (50 male speakers). We set the parameter values δ as 8.0 and β as 0.01. Both of the initial HMM training and ODT iteration, reduction of recognition error rates are shown, changing the training data set from DST1 to DST2.

Table 1 Recognition result.

Training data set		Error rate(%)		Error reduction(%)
for MLE	for ODT	MLE	ODT	
DST1(10male)	DST1	3.9	1.7	55.2
DST2(50male)	DST1	2.0	1.4	30.6
DST2	DST2	2.0	0.8	55.1

Iteration counts The near-miss rate on condition $D(c,n) < 0$ in Figure 4 is equivalent to the the error rates for training data. When DST1 are used for initial HMM training and ODT iteration, the HMMs strongly depend on each speaker of the training data, so the curve of the recognition error rate with the iteration is quite different from the curve of the error rate for training data. When DST2 are used for initial HMM training, these curves show a similar tendency. It seems that the dependency on each speaker of the training data reduces by increasing the training members for the HMMs. So, if we terminate the ODT iteration, when the error rate of training data becomes 0 or the error-rate-decrement with the iteration becomes small enough, then we can obtain the HMMs of better performance without recognition experiments.

3.3 Results and discussion

Table 1 shows the recognition results corresponding to Figure 4, comparing the MLE method with the ODT method. As continuous mixture density HMMs are used for this recognition system, so even with the MLE method, good recognition performance can be obtained. And, the ODT method makes a significant improvement of the performance of the HMMs. In these experiments, speakers of the data is different from that of the test data, but the results show very high recognition accuracy of the HMMs. Especially, by using DST2 of 50 speakers for initial HMM training and the ODT iteration, final recognition error rate is reduced to only 0.8%.

The recognition results, when DST1 of 10 speakers is used for the ODT iteration, show that the performance of initial HMM can be improved by small number of training set with the ODT method. This result leads that the ODT method is applicable to the adaptation of HMMs to some vocabulary, by using small number of word data. We are investigating a speaker-independent word recognition system based on sub-phonetic unit HMMs, that the recognition vocabulary can be registered without a large amount of training word data. We expect that, by using the ODT method, the recognition performance of the system can be improved with a small amount of the registered word data.

4. CONCLUSION

In this paper, we described the ODT method for continuous mixture density HMMs, considering the recognition performance of HMMs in the training. We applied this method to speaker-independent isolated word recognition.

We discussed the contribution of each parameter of the ODT method to the recognition performance, and the way to optimize the parameters without recognition experiments. We summarize as follows: when the near-miss criterion δ were large enough so that most of the training data is used in the ODT iteration, and a value of learning rate $\beta$was small enough, the HMMs obtained the best recognition performance. We can determine the value of these parameters by above mentioned way, without recognition experiments. By increasing the number of speakers, both for initial HMM training and ODT iteration, the recognition performance are improved. And on the other hand, even if the number of speakers for the ODT iteration was small, some improvement was obtained. When the initial HMM are trained from large number of speakers, the error rate of training data and recognition error rate showed similar tendency. So, we can determine the iteration counts of the ODT, from the error number of training data.

Finally, the significant effectiveness of this method were shown by the results of the recognition experiments.

References

[1] L.R.Rabinar, B-H.Juang, S.E.Levinson, M.M.Soundhi: "Recognition of isolated digits using hidden Markov models with continuous mixture densities", AT&T. Tech. J., 64, 6, pp.1211-1231 (1985).

[2] L.R.Bahl, P.F.Brown, P.V.de Sauza, R.L.Mercer: "A new algorithm for the estimation of HMM parameters", Proc. IEEE ICASSP, S11.2, New York(1988).

[3] Kai-Fu Lee, S.Mahajan: "Corrective and reinforcement learning for speaker-independent continuous speech recognition", Technical Report CMU-CS-89-100, Carnegy Mellon University (Jan.1989).

[4] T.H.Applebaum, B.A.Hanson: "Enhancing the discrimination of speaker independent hidden Markov models with corrective training", Proc. IEEE ICASSP, S6.13,Glasgow(1989).

[5] S.Mizuta, K.Nakajima: "Optimum discriminative training for HMM with continuous mixture densities", Proc. Spring Meet. Acoust.Soc.Japan 1-3-12 (1990) (in Japanese).

[6] T.Kohnen, G.Banra, R. Chrisley: "Statistical pattern recognition with neural networks", IEEE, Proc. of ICNN, Vol. I, pp.61-68 (Jul.1988).

[7] S.Furui: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans., Acoust., Speech, Signal Processing, ASSP-34, 1, pp.52-59(1986).

[8] S.E.Levinson: "Continuously variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language, 1, pp.29-45 (1986).