A discriminative training method for continuous mixture density HMMs and its implementation to recognize noisy speech

Sinobu Mizuta and Kunio Nakajima

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

(Received 21 February 1992)

In this paper, a training method for continuous mixture density HMMs, named optimal discriminative training (ODT), and its implementation for speech recognition in noise are described. ODT is one of corrective learning methods, applied to continuous mixture density HMMs, and these HMMs are especially useful for speaker-independent speech recognition. Under noisy environments, the recognition categories are liable to confuse, so by using ODT the improvement of recognition accuracy is more expected. Here, we describe the training algorithm of ODT, and the effects of ODT to improve the robustness for adverse environments by the word recognition experiments in noise.

Keywords: Speech recognition, Hidden Markov model, Optimal discriminative training, Noisy speech

PACS number: 43.72. Ne

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.\(^1\) 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. Particularly, the latter is effectable for speaker-independent recognition where the varied acoustic features must be treated.

Generally, training methods for HMM parameters are based on the maximum likelihood estimates (MLE), but MLE has a problem that the discriminative accuracy between the confusable categories is not considered in the training. To solve the problem for continuous mixture density HMMs, we have already proposed an error corrective training method called the optimal discriminative training (ODT). In this method, some parameters of each HMMs are modified to minimize the misrecognition and near-misrecognition for training data set, so under the training condition, optimal discrimination for these data can be obtained. When using ODT, the improvement of recognition performance has been shown by some experiments.\(^2\)

Under noisy environments, the distance between data sets of different classes is close in pattern space, and the discrimination of these data becomes difficult. In training methods based on MLE, the confusion of classes is not considered, so the methods cannot reduce the influence of the problem sufficiently. Therefore, the improvement of recognition accuracy by using error corrective training, such as ODT, is more expected.

In this paper, firstly we describe the training procedure of ODT. Then, we investigate the effects of
ODT for noisy speech, from speaker-independent word recognition experiments.

2. OPTIMAL DISCRIMINATIVE TRAINING

Figure 1 shows a block diagram of the recognition system using continuous mixture density HMMs based on ODT. In ODT, each sample of the training data set is recognized using initial HMMs. These HMMs are trained by MLE. 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 HMMs 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 paths 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 Gaussian mixture densities.
6. Move the vectors \( \mu(c, i), \mu(n, i) \) of \( m(c, i), m(n, i) \) by following way:

\[
\begin{align*}
\mu(c, i) &= \mu(c, i) + \gamma(V(i) - \mu(c, i)) \\
\mu(n, i) &= \mu(n, i) - \gamma(V(i) - \mu(n, i))
\end{align*}
\tag{2}
\]

Where,

\[
\gamma = \begin{cases} 
\frac{\beta}{(D(c, n) - \delta)} & (D(c, n) > \delta) \\
\beta & (D(c, n) < 0) \\
\beta(1 - D(c, n)/\delta) & (0 \leq D(c, n) \leq \delta)
\end{cases}
\tag{4}
\]

\[
\beta > 0, \delta > 0
\tag{5}
\]

7. Continue with step 2, until the number of iterations have been large enough to minimize the number of misrecognition and near-misrecognition for the training data set.

Figure 2 is an illustration about the movement of the mean vectors \( \mu(c, i) \) and \( \mu(n, i) \). In the feature space, the training sample \( L(c) \) is represented as a locus, and the HMMs \( H(c) \) and \( H(n) \) can also be represented statistical locuses. After the correspondence between the training sample and HMMs is obtained on step 4, \( \mu(c, i) \) and \( \mu(n, i) \) are modified with correspondent frame vector \( V(i) \) on step 6. The modification of these vectors is achieved by similar manner to the means in LVQ2.4)

In the preceding procedure, learning rate \( \beta \) and near-miss criterion \( \delta \) are parameters having strong effects on final recognition accuracy of trained HMMs. Figure 3 shows the relation between \( D(c, n) \)
Fig. 3 Relation among training parameters.

and $\gamma$, where $D(c, n)$ is the difference between the probabilities, and $\gamma$ is the final moving rate. In this figure, $\delta$ and $\beta$ are represented as the intercepts of each axis. We have investigated about the optimization of these parameters.\(^5\)\(^6\) After the optimization, the significant effectiveness of ODT can be shown.

3. EFFECTS OF ODT FOR NOISY SPEECH

Recently, various kinds of speech recognition system are proposed, but generally they are not provided the robustness to noisy signal. A main factor of this problem is that, in noisy condition, the distance between speech data of the different categories is close in the pattern space. In the HMM training by MLE, the discriminative accuracy between the categories is not considered, so the recognition system using the HMMs trained by this method cannot avoid the influence of the problem.

ODT is a training method based on an evaluation of recognition accuracy for training data, so the following effects can be expected.

1. If the tendency of recognition error under noisy environment is similar to that under quiet environment, then, when applying ODT to clean training data, the recognition performance raises both for clean and noisy data.

2. When applying ODT to noisy training data directly, the recognition performance raises for noisy data. If the Signal-to-Noise Ratio (SNR) of the training data is various, the robustness for noisy environment can be obtained.

In following section, we discuss these effects of ODT in some experiments of speaker-independent word recognition.

4. EXPERIMENTAL EVALUATION

4.1 Speech Representation

Speech signals are sampled at 8 kHz, and pre-emphasized with a filter of $1 - 0.95z^{-1}$. Then, a Hamming window with a width of 25.6 ms is applied every 10 ms, and 10 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,\(^5\) and regression coefficient of the mean energy.

Word HMMs for the recognition experiments are comprised of sub-phonetic unit HMMs. As the sub-phonetic unit HMMs, the duration controllable HMMs\(^6\) 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 50 male speakers and 50 female speakers are used. The data set of 25 male speakers and 25 female speakers with SNR $\infty$ and 10 db (DS1[$\infty$], DS1[10]) is used as training data, and the data set of 25 male speakers and 25 female speakers, different from DS1, with SNR $\infty$, 20, and 10 db (SD2[$\infty$], DS2[20], DS2[10]) is used as test data. Noisy data are generated artificially by adding white noise to clean speech.

In the ODT iteration and the recognition, we selected only 10 confusable words for each sample, to reduce the computation time. Learning rate $\beta$ and near-miss criterion $\delta$ on ODT determined as 0.01 and 2.0.

4.2 Effects of ODT Using Clean Data

Figure 4 (dotted line) shows the effects of ODT for various noisy speech, using DS1[$\infty$] for both MLE and ODT. Where, the error rate for DS2[$\infty$] decreases with the ODT iteration, but on the contrary, that for DS2[10] increases.

Table 1 shows the examples of recognition error for initial HMMs. This table indicates that the tendency of recognition error under noisy environment is largely different from the tendency under quiet environment, and this result is contrary to the expectation in section 3. It seems that, the category sets separated by ODT using clean data are different from that the category sets confusing each other in

N. MIZUTA and K. NAKAJIMA: DISCRIMINATIVE TRAINING METHOD FOR HMMS

Fig. 3 Relation among training parameters.
4.3 Effects of ODT Using Noisy Data

Figure 4 (dashed line) shows the effects of ODT for various noisy speech, with DS1[∞] for MLE, and DS1[10] for ODT. When comparing with the effects of ODT using only DS1[10], high recognition performance for DS2[10] is similarly obtained, at the same time, the increase of error rate for DS2[∞] occurred on the previous experiment is not seen here. This result indicates that, by ODT using various noisy speech, the robustness for noisy environment can be obtained.

4.4 Improvement of the Robustness for Noise

Figure 4 (solid line) shows the effects of ODT for various noisy speech, using DS1[∞] for MLE, and both DS1[∞] and DS1[10] for ODT. When comparing with the effects of ODT using only DS1[10], high recognition performance for DS2[10] is similarly obtained, at the same time, the increase of error rate for DS2[∞] occurred on the previous experiment is not seen here. This result indicates that, by ODT using various noisy speech, the robustness for noisy environment can be obtained.

4.5 Influence of Initial HMM Training

Figure 5 shows the effects of ODT for various noisy speech, using both DS1[∞] and DS1[10] for ODT, and various training data set for MLE (only DS1[∞], only DS1[10], and both of them). In these three cases, when using both DS1[∞] and DS1[10] for MLE, the highest recognition performance is obtained for every test data. And even in other cases high performances are obtained similarly, and these results indicates that the influence of initial HMMs is comparatively small.

This figure also shows, to compare the previous results, the effects of ODT for DS2[∞], using DS1[∞] for both ODT and MLE, and the effects for DS2[10], using DS1[10] for these trainings.

noisy speech recognition, and this difference causes the phenomena in Fig. 4.

Table 1 Examples of recognition error for initial HMMs.

<table>
<thead>
<tr>
<th>Error/Input (SNR)</th>
<th>∞</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>meguro</td>
<td>→ nemuro</td>
<td>10</td>
</tr>
<tr>
<td>kiriju</td>
<td>→ ciruju</td>
<td>9</td>
</tr>
<tr>
<td>takeo</td>
<td>→ ageo</td>
<td>6</td>
</tr>
<tr>
<td>noDa</td>
<td>→ mooKa</td>
<td>2</td>
</tr>
<tr>
<td>hoNDNdo</td>
<td>→ hoNjoo</td>
<td>2</td>
</tr>
<tr>
<td>ciruju</td>
<td>→ kiruju</td>
<td>2</td>
</tr>
<tr>
<td>oobu</td>
<td>→ koobu</td>
<td>1</td>
</tr>
<tr>
<td>gotEnba</td>
<td>→ keseNnuma</td>
<td>0</td>
</tr>
<tr>
<td>ageo</td>
<td>→ takeo</td>
<td>0</td>
</tr>
<tr>
<td>bisai</td>
<td>→ hisai</td>
<td>0</td>
</tr>
<tr>
<td>warabi</td>
<td>→ asahi</td>
<td>0</td>
</tr>
<tr>
<td>juuki</td>
<td>→ zushi</td>
<td>0</td>
</tr>
<tr>
<td>kitami</td>
<td>→ asahi</td>
<td>0</td>
</tr>
</tbody>
</table>

This figure also shows, to compare the previous results, the effects of ODT for DS2[∞], using DS1[∞] for both ODT and MLE, and the effects for DS2[10], using DS1[10] for these trainings.
Table 2 Recognition results.

<table>
<thead>
<tr>
<th>Training data set (SNR)</th>
<th>Iteration</th>
<th>Error rate of various SNR input (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>∞</td>
</tr>
<tr>
<td>for MLE</td>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>for ODT</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>∞</td>
<td>∞</td>
<td>45.6</td>
</tr>
<tr>
<td>10 dB</td>
<td>10 dB</td>
<td>5</td>
</tr>
<tr>
<td>∞ + 10 dB</td>
<td>∞ + 10 dB</td>
<td>1.4</td>
</tr>
<tr>
<td>∞ + 10 dB</td>
<td>∞ + 10 dB</td>
<td>0.8</td>
</tr>
</tbody>
</table>

When using both DS1[∞] and DS1[10] jointly for both training process, it seems to be obtained the similar recognition performance to the case that is dependent on the noise condition of the test data. These final recognition results are shown on Table 2.

5. CONCLUSION

In this paper, we described the effects of ODT, a training method for continuous mixture density HMMs considering recognition accuracy, for noisy speech. We investigated these effects from speaker-independent isolated word recognition experiments.

When using noisy data for ODT, recognition performance for noisy speech was improved greatly, and the undesirable influence for clean speech was very few. When the various noise condition were applied for ODT training data, the robustness for noisy environment was obtained.

When using these various noisy speech also for MLE, the highest performance was obtained, but the initial HMM training data set for MLE had few influence.

As a result, ODT for HMMs lead to the great improvement of recognition accuracy for noisy speech on the recognition experiments.

REFERENCES


Shinobu Mizuta received the B. S. degree in Information Science from Kyoto University, Kyoto, Japan in 1988. He is currently working on the development of speech recognition technology at the Computer and Information Systems Laboratory, Mitsubishi Electric Corporation.

Kunio Nakajima received the B. S. and M. S. degrees in electronic engineering from Nagoya Institute of Technology, Nagoya, Japan in 1968 and 1970, respectively. Since 1970, he has been working at Mitsubishi Electric Corporation, and he is currently the manager of the Speech Processing Group in the Computer and Information Systems Laboratory. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan, and the Society of Information Theory and Its Applications.