Rapid EM Training based on Model-Integration

Shinichi Yoshizawa +, Kiyohiro Shikano ++

+ Matsushita Electric Industrial Co., Ltd., Japan, ++ Nara Institute of Science and Technology, Japan
+ yoshizawa.shinichi@jp.panasonic.com, ++ shikano@is.aist-nara.ac.jp

Abstract

Recently, speech recognition technique has started being used in various products. In order to make a good acoustic model, usually a lot of training speech data is needed. However, due to the right of voice and privacy issues, it is not easy to collect a lot of training data. Statistical models which have been trained and transformed from speech data do not have the above mentioned problem, and are comparatively easy to obtain. Therefore, a training technique which can make an acoustic model by using statistical models as training data is preferred. This paper describes a new method called Rapid EM Training based on Model-Integration and shows that the proposed method is evaluated positively.

1. Introduction

The Baum-Welch algorithm [1] has been widely used for training an acoustic model. This method calculates an acoustic model by using a lot of training speech data, and can obtain a high recognition rate. However, due to the right of voice and privacy issues, collecting speech data costs a lot and it is not easy to collect enough training data.

On the other hand, statistical models which have been trained and transformed from speech data do not have the above mentioned problem, and are relatively easy to obtain through the Internet. Some statistical models can be collected from speech organizations’ website. Therefore, a training technique based on statistical models is expected.

This paper describes a new method called Rapid EM (Expectation-Maximization) Training based on Model-Integration [2]. Our proposed method estimates an acoustic model by using only statistical models as training data and obtains good experimental results for a various kinds of conditions. It utilizes maximum likelihood criterion same as the Baum-Welch [1], and can get high recognition rate equal to the Baum-Welch. Additionally, the rapid training is executed by using a small number of statistics of training statistical models and can get an acoustic model 100 times faster than the Baum-Welch.

2. EM training based on statistical models

The proposed EM training is described in Fig.1. The procedure of this method is the same as the Baum-Welch, but statistical models are used as training data instead of speech data. First, statistical models are prepared, and these models consist of hidden Markov models (HMMs) which have continuous output density of mixture of Gaussian distributions. Then, an initial acoustic model \( \{\omega_f(m) | \mu_f(m), \sigma^2_f(m) \} \) \( (\omega_f(m) | \mu_f(m), \sigma^2_f(m) \) in sec.2.2) is set up. Finally, the acoustic model is updated from the initial model by using only statistical models on the basis of maximum likelihood criterion.

2.1. Optimization function

The proposed method’s optimization function is log-likelihood for the training statistical models as follows:

\[
\log P = \sum_{i=1}^{N_x} \int_{-\infty}^{\infty} \{ \log [ \sum_{m=1}^{M_f} \omega_f(m) f(x; \mu_f(m), \sigma^2_f(m)) ] \\
\sum_{i=1}^{I_g(l)} \omega_g(l,i) g(x; \mu_g(l,i), \sigma^2_g(l,i)) \} dx \tag{1}
\]
where \( N_g \) is the number of training models, and \( f(\cdot) \) and \( g(\cdot) \) are Gaussian distributions of estimated acoustic model and of training models, respectively. \( M_f \) and \( L_{g(i)} \) are the number of mixture of Gaussian of estimated acoustic model and of \( i \)th training model, respectively. \( \omega_f(\cdot) \) and \( \omega_g(\cdot) \) are weights, \( \mu_f(\cdot) \) and \( \mu_g(\cdot) \) are means, and \( \sigma^2_f(\cdot) \) and \( \sigma^2_g(\cdot) \) are variances, of \( m \)th Gaussian of estimated acoustic model and of \( l \)th Gaussian of \( i \)th training model, respectively.

Statistics of an acoustic model (\( \omega_f(\cdot) \), \( \mu_f(\cdot) \), and \( \sigma^2_f(\cdot) \)) in each state of HMM are repeatedly calculated as follows:

### 2.2. Estimation of an acoustic model

To maximize the optimization function, statistics of an acoustic model are repeatedly calculated as follows:

\[
\omega_f(m)[t+1] = \frac{\sum_{i=1}^{N_g} A(m,i)[t]}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} A(k,i)[t]} \quad (m=1,2,\ldots,M_f) \tag{2}
\]

\[
\mu_f(m,j)[t+1] = \frac{\sum_{i=1}^{N_g} B(m,i,j)[t]}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} A(m,i)[t]} \quad (m=1,2,\ldots,M_f, j=1,2,\ldots,J) \tag{3}
\]

\[
\sigma^2_f(m,j)[t+1] = \frac{\sum_{i=1}^{N_g} C(m,i,j)[t]}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} A(m,i)[t]} \quad (m=1,2,\ldots,M_f, j=1,2,\ldots,J) \tag{4}
\]

where \( t \) is EM iteration times and \( j \) is index of dimension of \( x \), and \( A(m,i)[t] \), \( B(m,i,j)[t] \), and \( C(m,i,j)[t] \) are calculated as follows:

\[
A(m,i)[t] = \int_{-\infty}^{\infty} \{ \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i) g(x;\mu_g(l,i),\sigma^2_g(l,i)) \} dx \tag{5}
\]

\[
B(m,i,j)[t] = \int_{-\infty}^{\infty} x_j \{ \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i) g(x;\mu_g(l,i),\sigma^2_g(l,i)) \} dx \tag{6}
\]

\[
C(m,i,j)[t] = \int_{-\infty}^{\infty} (x_j - \mu_f(m,j)[t])^2 \times \{ \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i) g(x;\mu_g(l,i),\sigma^2_g(l,i)) \} dx \tag{7}
\]

where

\[
\gamma(x;m)[t] = \frac{\omega_f(m)[t] f(x;\mu_f(m)[t],\sigma^2_f(m)[t])}{\sum_{k=1}^{M_f} \omega_f(k)[t] f(x;\mu_f(k)[t],\sigma^2_f(k)[t])} \tag{8}
\]

Transition probabilities of HMM of an acoustic model are calculated as follows:

\[
T_f(i)[j] = \frac{\sum_{k=1}^{M_f} \sum_{l=1}^{N_g} \gamma(x;m)[t] T_f(k)[l] \omega_g(l,i)[t]}{\sum_{j=1}^{N_g} \sum_{i=1}^{M_f} \gamma(x;m)[t] \omega_g(l,i)[t]} \tag{9}
\]

where \( T_f(i)[j] \) and \( T_g(l)[k][j] \) are transition probabilities from \( i \)th state to \( j \)th state of HMM of estimated acoustic model and \( l \) and of \( k \)th training model, respectively, and \( N_{st} \) is the number of states.

### 2.3. Rapid EM training

Gaussians of an estimated acoustic model are assumed to be hardly overlapped with each other [2], and that leads to the approximation of \( \gamma(x;m)[t] \) in eq.(8) as follows:

\[
\gamma(x;m)[t] \approx \begin{cases} 1 & \text{for } \omega_f(m)[t] \text{ nearby } f(m)[t] \\ 0 & \text{otherwise} \end{cases} \tag{10}
\]

Using this approximation, the rapid estimation formula is obtained as follows:

\[
\omega_f(m)[t+1] \approx \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t]}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t]} \tag{11}
\]

\[
\mu_f(m,j)[t+1] \approx \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t] \mu_g(l,i,j)}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t]} \tag{12}
\]

\[
\sigma^2_f(m,j)[t+1] \approx \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t] \mu_g(l,i,j) + \mu^2_g(l,i,j)}{\sum_{i=1}^{M_f} \sum_{j=1}^{N_g} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)[t] \omega_g(l,i)[t]} - \mu^2_f(m,j)[t+1] \tag{13}
\]

where \( \gamma(x;m)[t] \) is in eq.(10).

The rapid training can be executed by using a small number of statistics of training models.

### 3. Experimental results

The proposed method is investigated to obtain high recognition rate by using only statistical modes, and to execute the rapid training compared to the Baum-Welch algorithm [1].

Japanese speech corpus collected by Acoustical Society of Japan [3] is used in our experiments. This database
3.1. Noisy condition’s acoustic model

An acoustic model which can work on every 6 kinds of noisy environment (vacuum cleaner, department store, train, platform, car, intersection) is obtained. Monophone HMM with 16-mixture is used for both an acoustic model and training models, and 6 training models are prepared which have been trained by using speech data on 6 kinds of noisy environment. The Baum-Welch uses training speech data which is used for making training models. Clean model is used for an initial acoustic model.

The results of the proposed method and the Baum-Welch are shown in Table 1, and word accuracy and training time are described. Word accuracy of 55.1% is obtained before training. From Table 1, the proposed method obtains high word accuracy of 59.9% similar to the Baum-Welch, showing EM training for statistical models method works well. The word accuracy of the proposed method is slightly worse than the Baum-Welch, and one of the reason might be that the ill-posed problem occurs because only 6 training models are used. As for the training time, the proposed rapid training method takes only 8 seconds for execution, compared to 11 hours needed for the Baum-Welch.

3.2. Speaker adapted acoustic model

Speaker adapted acoustic model is obtained by using training models which are acoustically close to test speaker. Training models are selected by using likelihood for testing data (as refer in [5]). Monophone HMM with 16 or 64-mixture is used for an acoustic model and training models, and 10 or 20 training models are prepared. All speakers’ model which is trained from 260 training speakers is used for an initial acoustic model.

The result for an acoustic model with 16-mixture is shown in Fig. 2, and the one for an acoustic model with 64-mixture is shown in Fig. 3. An acoustic model is trained by using the same mixture’s training models. Word accuracy of 82.2% (16-mixture) or 86.3% (64-mixture) is obtained before training. From Fig. 2 and
Figure 3: **Word accuracy and training time**: speaker adapted acoustic model with 64-mixture from training models with 64-mixture.

The proposed method obtains high word accuracy of 85.7% (16-mixture) or 89.6% (64-mixture) similar to the Baum-Welch. The word accuracy of the Baum-Welch is decreased as the iteration time increases, and the reason is that over-fitting might have occurred. As for the training time, the proposed method takes 12 seconds (16-mixture) or 6 minutes (64-mixture), and the rapid training is executed much faster than the Baum-Welch which takes 25 minutes (16-mixture) or 2.6 hours (64-mixture).

The result for an acoustic model with 16-mixture from training models with 64 is shown in Fig.4. An initial acoustic model is made by selecting 16 Gaussian randomly from all speaker models with 64-mixture. Word accuracy of 74.7% is obtained before training. From Fig.4, the proposed method obtains high word accuracy of 85.7% (after 5 times iteration) similar to the Baum-Welch, while the training time of the proposed method is quite faster than the Baum-Welch.

### 4. Conclusions

A new type of EM training is proposed, and can make an acoustic model by using only statistical models. The proposed method can obtain high recognition rate equal to or better than the Baum-Welch algorithm. Additionally, the rapid training is executed and can get an acoustic model 100 times faster than the Baum-Welch. In our future work, we will investigate the methods of making an initial acoustic model, and will try to overcome the ill-problem.

### 5. References


