

Model-Integration Rapid Training based on Maximum Likelihood for Speech Recognition

Shinichi Yoshizawa ⁺, Kiyohiro Shikano ⁺⁺

⁺ Matsushita Electric Industrial Co.,Ltd.,Japan,
⁺⁺ Nara Institute of Science and Technology, Japan
⁺ yoshizawa.shinichi@jp.panasonic.com

Abstract

Speech recognition technology has been widely used. Considering a training cost of an acoustic model, it is beneficial to reuse pre-existing acoustic models for making a suitable one for various apparatus and application. However, a complex acoustic model for high CPU power does not work for low CPU power. And a simple model for fast-processing-demanded application does not work well for high-precision-demanded ones. Therefore, it is important to adjust a model complexity according to apparatus or application, such as a number of mixture of Gaussians. This paper describes a new model-integration-type of training for obtaining a required number of mixture of Gaussians. This training can alter a number of mixture into a required one according to a specification of apparatus or application. We propose a model integration rapid training based on maximum likelihood, and evaluate the recognition performance successfully.

1. Introduction

Various kinds of acoustic model's trainings have been proposed. Baum-Welch algorithm trains an acoustic model using speakers' utterance data based on maximum likelihood. This training can obtain a required number of mixture of Gaussians, and it can also obtain a high-precision acoustic model. However, this training uses a lot of utterance data and it takes a lot of time to make an acoustic model.

To solve this problem, model-integration-type of trainings have been proposed. "By Sufficient Statistics Speaker Adaptation"[1] and "Cluster Adaptive Training"[2] are the type of this training. These trainings can execute a rapid training because an acoustic model is calculated by a small number of parameters of pre-existing acoustic models. However, these trainings are quite difficult to make a required number of mixture of Gaussians because the number is restricted to the one of pre-existing acoustic models.

This paper describes a new model-integration-type of training for obtaining a required number of mixture of Gaussians. The proposed training can alter a number of mixture into a required one according to a specification of apparatus or application. Besides, a rapid training can be executed by using a small number of statistics of pre-existing acoustic models, and this training can obtain a high-precision model because of a training based on maximum likelihood. When pre-existing models have been trained by data acoustically close to specific speakers, speaker adaptation can also be executed.

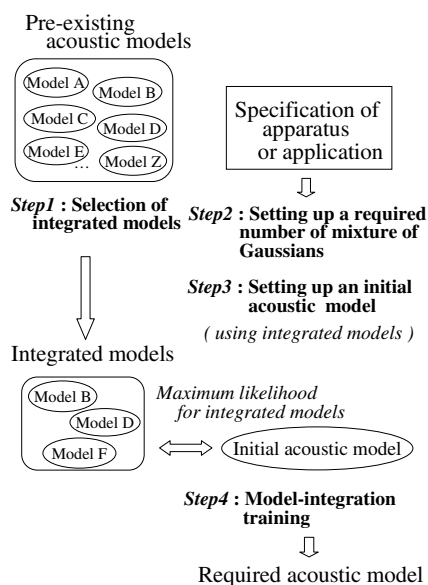


Figure 1: Block diagram of the proposed training.

2. Model-integration training based on maximum likelihood

The proposed model-integration training is described in Fig.1. The training consists of 4 steps. In the first step, integrated models are selected from pre-existing acoustic models. An acoustic model is calculated by using the integrated models. In the second step, a required number of mixture of Gaussians is set up. The number is decided according to a specification of apparatus or application. In the third step, an initial acoustic model is set up. In the final step, model-integration training based on maximum likelihood is executed by using the integrated models from the initial acoustic model.

2.1. Selection of integrated models

In the first step, integrated models are selected from pre-existing acoustic models. Pre-existing acoustic models consist of hidden Markov models (HMMs) which have continuous output density of mixture of Gaussian distributions, and they have a various number of mixture.

When integrated models have been trained by data acousti-

cally close to specific speakers, speaker adaptation can also be executed. The integrated models can be selected as in the same way of ‘‘By Sufficient Statistics Speaker Adaptation’’[1].

2.2. Setting up a required number of mixture of Gaussians

In the second step, a required number of mixture of Gaussians is set up. The proposed training can decide a number of mixture (M_f in sec.2.4.1) independent of the one of the integrated models according to a specification of apparatus or application. While, traditional model-integration-type of trainings are difficult to set up a required number of mixture because the number is restricted to the one of the integrated models.

2.3. Setting up an initial acoustic model

In the third step, an initial acoustic model ($\omega_{f(m)[0]}$, $\mu_{f(m,j)[0]}$, and $\sigma_{f(m,j)[0]}$ in sec.2.4.2) is set up. The initial model is quite important to obtain a high-precision acoustic model. This paper describes a method to make an initial model from the integrated models.

2.4. Model-integration training

In the final step, a new model-integration-type of training is proposed. An acoustic model is estimated from the initial acoustic model based on maximum likelihood for the integrated models. A rapid training can be executed by directly using statistics of the integrated models. This training procedure is the same as EM algorithm.

2.4.1. Evaluation function

Evaluation function is a likelihood for the integrated models not for training utterance data. Statistics of an acoustic model ($\omega_{f(m)}$, $\mu_{f(m)}$, and $\sigma_{f(m)}$) in a certain state of HMM are estimated by maximizing the evaluation function as follows:

$$\log L = \sum_{i=1}^{N_g} \int_{-\infty}^{\infty} \left\{ \log \left[\sum_{m=1}^{M_f} \omega_{f(m)} f(x; \mu_{f(m)}, \sigma_{f(m)}^2) \right] \right. \\ \left. \sum_{l=1}^{L_g(i)} v_{g(l,i)} g(x; \mu_{g(l,i)}, \sigma_{g(l,i)}^2) \right\} dx \quad (1)$$

where N_g is a number of integrated models, and $f(\cdot)$ and $g(\cdot)$ are Gaussian distributions of estimated acoustic model and of integrated models, respectively. M_f and $L_g(i)$ are a number of mixture of Gaussians of estimated acoustic model and of i th integrated model, respectively. $\omega_{f(m)}$ and $v_{g(l,i)}$ are weights, $\mu_{f(m)}$ and $\mu_{g(l,i)}$ are means, and $\sigma_{f(m)}^2$ and $\sigma_{g(l,i)}^2$ are variances, of m th Gaussian of estimated acoustic model and of l th Gaussian of i th integrated model, respectively,

2.4.2. Estimation of an acoustic model

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

$$\omega_{f(m)[t+1]} = \frac{\sum_{i=1}^{N_g} A(m, i)[t]}{\sum_{k=1}^{M_f} \sum_{i=1}^{N_g} A(k, i)[t]} \quad (2)$$

$$\mu_{f(m,j)[t+1]} = \frac{\sum_{i=1}^{N_g} B(m, i, j)[t]}{\sum_{i=1}^{N_g} A(m, i)[t]} \quad (3)$$

$$\sigma_{f(m,j)[t+1]}^2 = \frac{\sum_{i=1}^{N_g} C(m, i, j)[t]}{\sum_{i=1}^{N_g} A(m, i)[t]} \quad (4)$$

where t is 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:

$$A(m, i)[t] = \int_{-\infty}^{\infty} \gamma(x; m)[t] \left\{ \sum_{l=1}^{L_g(i)} v_{g(l,i)} g(x; \mu_{g(l,i)}, \sigma_{g(l,i)}^2) \right\} dx \quad (5)$$

$$B(m, i, j)[t] = \int_{-\infty}^{\infty} \gamma(x; m)[t] x_j \left\{ \sum_{l=1}^{L_g(i)} v_{g(l,i)} g(x; \mu_{g(l,i)}, \sigma_{g(l,i)}^2) \right\} dx \quad (6)$$

$$C(m, i, j)[t] = \int_{-\infty}^{\infty} \gamma(x; m)[t] (x_j - \mu_{f(m,j)[t]})^2 \left\{ \sum_{l=1}^{L_g(i)} v_{g(l,i)} g(x; \mu_{g(l,i)}, \sigma_{g(l,i)}^2) \right\} dx \quad (7)$$

where

$$\gamma(x; m)[t] = \frac{\omega_{f(m)[t]} f(x; \mu_{f(m)[t]}, \sigma_{f(m)[t]}^2)}{\sum_{k=1}^{M_f} \omega_{f(k)[t]} f(x; \mu_{f(k)[t]}, \sigma_{f(k)[t]}^2)} \quad (8)$$

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

$$T_f[i][j] = \frac{\sum_{k=1}^{N_g} T_{g(k)}[i][j]}{\sum_{j=1}^{N_{st}} \sum_{k=1}^{N_g} T_{g(k)}[i][j]} \quad (9)$$

where $T_f[i][j]$ and $T_{g(k)}[i][j]$ are transition probabilities from i th state to j th state of HMM of estimated acoustic model and of k th integrated model, respectively, and N_{st} is a number of states.

2.4.3. Rapid training technique

The average overlapping of the nearest Gaussians for an acoustic model with 16 Gaussians is described above in Fig.2. The average of Mahalanobis generalized distance is 3.3. Here, we assume that Gaussians of an acoustic model are hardly overlapped each other.

On the assumption, $\gamma(x; m)$ in eq.(8) becomes quite simple as follows:

$$\gamma(x; m) = \begin{cases} 1 & v_{g(\cdot)} g(\cdot) \text{ nearby } \omega_{f(m)} f(x; \mu_{f(m)}, \sigma_{f(m)}^2) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

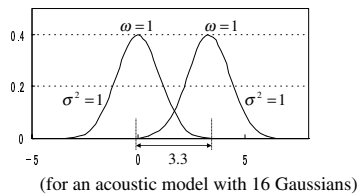
In Fig.2, the way of deciding of values of $\gamma(x; m)$ for the integrated models are described.

Substituting equation (10) for the equation (5), (6), (7), a new estimation formula is obtained as follows:

$$\omega_{f(m)[t+1]} = \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(m,i)}} v_{g(l,i)}}{\sum_{k=1}^{M_f} \sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(k,i)}} v_{g(l,i)}} \quad (11)$$

$$\mu_{f(m,j)[t+1]} = \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(m,i)}} v_{g(l,i)} \mu_{g(l,i,j)}}{\sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(m,i)}} v_{g(l,i)}} \quad (12)$$

The average overlapping of Gaussians



Estimated acoustic model

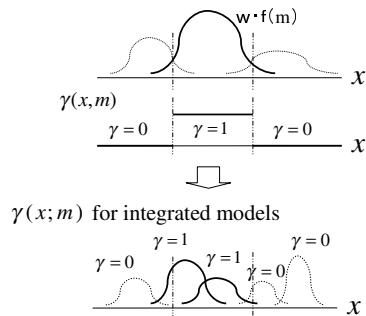


Figure 2: $\gamma(x; m)$ for Gaussians of integrated models.

$$\sigma_{f(m,j)[t+1]}^2 = \frac{\sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(m,i)}} v_{g(l,i)} (\sigma_{g(l,i,j)}^2 + \mu_{g(l,i,j)}^2)}{\sum_{i=1}^{N_g} \sum_{l=1}^{Q_{g(m,i)}} v_{g(l,i)}} - \mu_{f(m,j)[t+1]}^2 \quad (13)$$

where $Q_{g(m,i)}$ is a number of Gaussians $v_{g(\cdot)}g(\cdot)$ of i th integrated model nearby $\omega_{f(m)}f(x; \mu_{f(m)}, \sigma_{f(m)}^2)$ of an estimated acoustic model.

A rapid training can be executed by directly estimating by a small number of statistics of the integrated models.

3. Experimental results and discussion

Japanese speech corpus collected by Acoustical Society of Japan[3] is used in our experiments. This database consists of 306 speakers and each speaker uttered about 150 sentences. Speech data are sampled at 16kHz and 16bits. Twelfth-order mel-frequency cepstrum coefficients (MFCC) are calculated every 10ms. The cepstrum differences (delta-MFCC) and delta-power are also used. Cepstrum mean normalization (CMN) is performed based on the whole utterance average.

As integrated models, monophone HMMs of 43 phones have 3 states and each state has a mixture of 16 or 64 Gaussians. And as an estimated acoustic model, monophone HMMs of 43 phones have 3 states and each state has a mixture of 16 Gaussians. 46 speakers' data (23 male and 23 female) are used for testing data, which are not included in training of integrated models. Word accuracy for gender-independent acoustic model and training time are investigated for various methods. Performance evaluation is carried out using the Japanese dictation system Julius[4] with the 20k newspaper article language model.

The baseline training by Baum-Welch algorithm shows the average word accuracy of 82.4% and training time of about 1 day using 1GHz CPU. It takes a lot of time to obtain an acoustic model.

Integrated models \Rightarrow Estimated model
16 Gaussians \Rightarrow 16 Gaussians

Comparison with various methods

method	word accuracy		training time	
	initial (t=0)	training (t=5)		
Proposed	initial (t=0)	training (t=5)	10 sec (t=1)	
	init: male	71.6%		82.2%
	init: female	70.7%		80.9%
Baum-Welch (baseline)	—	82.4%	1 day	
By Sufficient Statistics	—	82.2%	5 sec	

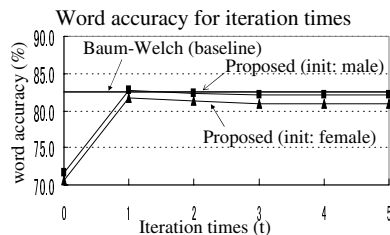


Figure 3: Word accuracy and training time for an acoustic model for required speakers.

3.1. Training for an acoustic model for required speakers

Training for obtaining an acoustic model for required speakers is shown. This experiment is a fundamental one for speaker adaptation. As integrated models, two monophone HMMs with 16 Gaussians for male and for female are selected. These models have been trained by Baum-Welch algorithm off-line. An acoustic model for gender-independent is estimated from the integrated models for male and for female by the proposed method. The integrated model for male or for female is used as an initial acoustic model.

In Fig.3, the results for the proposed method are described, and the results for “By Sufficient Statistics Speaker Adaptation”[1] are also described. The proposed method can obtain an acoustic model in about 10 seconds, and 1000 times faster than the training by Baum-Welch algorithm. And the proposed method can obtain high word accuracy of 82.2% (using male-based initial model) from 71.6% or 70.7% before training. Also, the results show that the way of making an initial model is quite important to obtain a high-precision acoustic model.

3.2. Training for an acoustic model with a required number of mixture of Gaussians

Training for obtaining an acoustic model with a required number of mixture of Gaussians is shown. This experiment is a fundamental one to obtain an acoustic model with a required number of mixture according to a specification of apparatus or application. Here, an acoustic model with 16 Gaussians is obtained from an integrated model with 64 Gaussians. As an integrated model, one monophone HMM with 64 Gaussians for gender-independent is selected. An acoustic model with 16 Gaussians for gender-independent is estimated from the integrated model by the proposed method. An initial acoustic model is made by picking up 16 Gaussians from the integrated model.

In Fig.4, the results for the proposed method are described. The proposed method can obtain a required number of mixture

Integrated model 64 Gaussians \Rightarrow Estimated model 16 Gaussians

Comparison with various methods

method	word accuracy		training time
	initial (t=0)	training (t=5)	
Proposed	init: gid	73.7%	10 sec (t=1)
		80.2%	
Baum-Welch (baseline)	—	82.4%	1 day

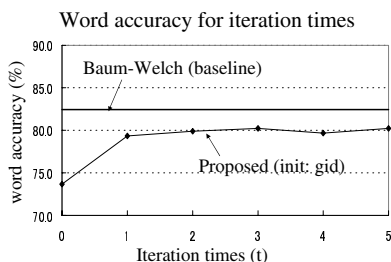


Figure 4: Word accuracy and training time for an acoustic model with a required number of mixture of Gaussians.

of Gaussians. The results show that the proposed method can obtain an acoustic model in about 10 seconds, and 1000 times faster than the training of Baum-Welch algorithm. And the proposed method can obtain high word accuracy of 80.2% from the word accuracy of 73.7% before training. While, “By Sufficient Statistics Speaker Adaptation”[1] can not obtain a required number of mixture’s model. To obtain a more high-precision acoustic model, the acoustic model obtained by the proposed method can be re-trained using a small number of sentence utterance data.

3.3. Training for an acoustic model for required speakers and with a required number of mixture of Gaussians

Training for obtaining an acoustic model for required speakers and with a required number of mixture of Gaussians is shown. Here, an acoustic model with 16 Gaussians is obtained from integrated models with 64 Gaussians. As integrated models, two monophone HMMs with 64 Gaussians for male and for female is selected. An acoustic model with 16 Gaussians for gender-independent is estimated from the integrated models by the proposed method. The initial acoustic model is made by selecting 16 Gaussians from the male-based integrated model or female-based one.

In Fig.5, the results for the proposed method are described. The proposed method can obtain a required number of mixture of Gaussians. The results show that the proposed method can obtain an acoustic model in about 20 seconds, and 1000 times faster than the training of Baum-Welch algorithm. And the proposed method can obtain high word accuracy of 80.4% from 65.2% or 63.3% before training. While, “By Sufficient Statistics Speaker Adaptation”[1] can not obtain a required number of mixture’s model.

Integrated models 64 Gaussians \Rightarrow Estimated model 16 Gaussians

Comparison with various methods

method	word accuracy		training time
	initial (t=0)	training (t=5)	
Proposed	init: male	65.2%	20 sec (t=1)
	init: female	63.3%	
		80.4%	
Baum-Welch (baseline)	—	82.4%	1 day

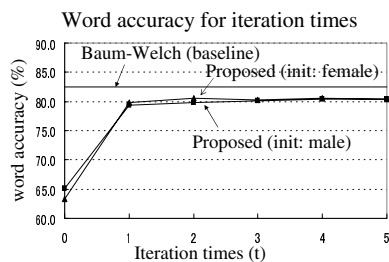


Figure 5: Word accuracy and training time for an acoustic model for required speakers and with a required number of mixture of Gaussians.

4. Conclusions

A new model-integration-type of training for obtaining a required number of mixture of Gaussians is proposed. The proposed training can alter a number of mixture of Gaussians into a required one according to a specification of apparatus or application. Besides, a rapid training can be executed by using pre-existing acoustic models. In future works, the way of making an initial acoustic model will be investigated, and an evaluation for speaker adaptation will be done.

5. References

- [1] Shinichi Yoshizawa, Akira Baba, Kanako Matsunami, Yuichiro Mera, Miichi Yamada, Kiyohiro Shikano, “Un-supervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers”, Proceedings of the ICASSP, 2001.
- [2] M.J.F. Gales, “Cluster Adaptive Training for Speech Recognition”, Proceedings of the ICSLP, pp.1783-1786, 1998.
- [3] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano and Shuichi Itahashi, ”JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research”, The Journal of the Acoustical Society of Japan (E), Vol.20, pp.199-206, 1999.
- [4] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda and Kiyohiro Shikano, ”A new phonetic tied-mixture model for efficient decoding”, Proceedings of the ICASSP, pp.1269-1272, 2000.