

論文 / 著書情報  
Article / Book Information

|           |                                                                             |
|-----------|-----------------------------------------------------------------------------|
| Title     | A Study of Speaker Adaptaion Based on Minimum Classification Error Training |
| Authors   | Tomoko Matsui, Sadaoki Furui                                                |
| Citation  | Eurospeech 95, Vol. 1, No. , pp. 81-84                                      |
| Pub. date | 1995, 9                                                                     |

# A STUDY OF SPEAKER ADAPTATION BASED ON MINIMUM CLASSIFICATION ERROR TRAINING

*Tomoko Matsui and Sadaoki Furui*

NTT Human Interface Laboratories  
3-9-11, Midori-Cho, Musashino-Shi, Tokyo, 180 Japan

## ABSTRACT

This paper studies a speaker adaptation method based on minimum-classification-error (MCE) training applied to a hidden Markov model (HMM) based speaker-independent speech recognition system. In this method, the HMM parameters are adapted to a new speaker using the combination of maximum a posteriori (MAP) and MCE estimation. MAP estimation maximizes the a posteriori probability that the HMMs generate the data of the speaker, but this does not always guarantee the highest performance for reducing the recognition error. On the other hand, MCE estimation directly aims at minimizing the recognition error and has recently started to be used for speaker adaptation. Here MCE estimation is applied to the HMM parameters adapted by MAP estimation so that they fall into one of the local minima near the HMM parameters adapted by MAP estimation.

In phoneme recognition experiments, we compare the performance of our combination of MAP and MCE estimation against the individual performances of MAP and MCE estimation. We find that the combination of MAP and MCE estimation is the most effective.

## 1 INTRODUCTION

Speaker adaptation is essential for HMM-based speech recognition systems, since the performance of speaker-independent phoneme HMMs for a new speaker is often low. Although there are many different approaches for speaker adaptation, the problem of insufficient amount of training data for the new speaker for estimating HMM parameters remains to be solved. When considering context-dependency of phoneme models, the number of HMM parameters becomes huge. To estimate them robustly using a small amount of data, some parameter-reduction techniques like parameter-tying are often used.

One frequently used method for adapting the HMM parameters to the speaker is maximum a posteriori (MAP) estimation [1][2]. It provides a way of incorporating prior information of the speaker-independent

system into the training process, which is particularly useful in dealing with problems posed by sparse training data for which the maximum-likelihood (ML) approach gives inaccurate estimates. However, in MAP estimation, there is no guarantee that the highest performance for reducing the recognition error will be obtained. On the other hand, minimum-classification-error (MCE) estimation aims at directly minimizing the recognition error and has recently started to be used for speaker adaptation [3]. One may regard MAP estimation as a method of estimating the feature parameter distribution for each phoneme accurately and MCE estimation as a method of adjusting the boundaries of the distributions effectively.

Therefore, a combination of methods might offer both advantages. For speaker recognition, Liu [5] has reported an effective method of making an HMM for each speaker that estimates the HMM parameters by maximum-likelihood (ML) estimation using the training data of the speaker and then modifies the parameters using MCE estimation.

This paper presents the combination of MAP and MCE estimation and compares it with MAP and MCE estimations and, in the discussion section, with the combination of ML and MCE estimation. For MCE estimation, the existence of many local minima is a serious problem especially when only a small amount of training data is available. We try to cope with this over-tuning problem in MCE estimation by estimating the model parameters so that they fall into one of the local minima near reasonable initial model parameters such as those adapted by MAP estimation. The combination improves the phoneme recognition performance most effectively, even when only one sentence is used as the training utterance for each speaker. In the experiments, we used tied-mixture HMMs as speaker-independent phoneme models because parameter-tying models are useful to cope with the problem of a small amount of data as stated above.

## 2 METHOD

In the combination of MAP and MCE estimation, the parameters of speaker-independent phoneme

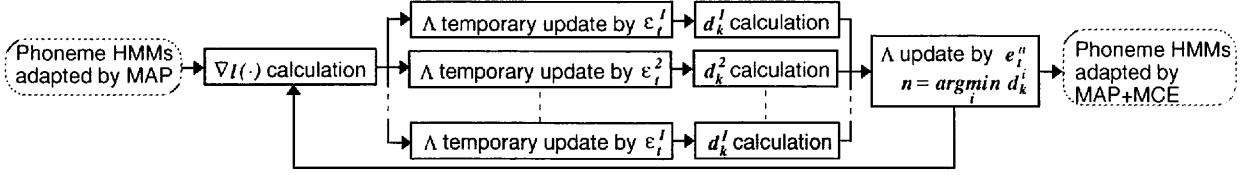


Figure 1. Procedure for adjusting  $\epsilon_t$

HMMs are adapted by MAP estimation and then modified using MCE estimation.

In MAP estimation [1][2], an HMM parameter set  $\Lambda$  is approximated while  $f(X|\Lambda)h(\Lambda)$  is locally maximized by the segmental MAP algorithm [4][1] where  $X$  is an observation sequence,  $\Lambda$  is an HMM parameter set,  $f(\cdot)$  is the likelihood function obtained by the Viterbi algorithm and  $h(\cdot)$  is the a priori density function.

The HMM parameters (a mean vector  $\mu_{sm}$  and weighting factor  $w_{sm}$  of a mixture component  $m$  in a state  $s$ ) are reestimated by eqs. (1) and (2).

$$\tilde{\mu}_{sm} = \frac{\tau_{sm}\mu_{sm} + \sum_{t=1}^T c_{smt}x_t}{\tau_{sm} + \sum_{t=1}^T c_{smt}} \quad (1)$$

$$\tilde{w}_{sm} = \frac{\nu_{sm} - 1 + \sum_{t=1}^T c_{smt}}{\sum_{m=1}^M \nu_{sm} - M + \sum_{m=1}^M \sum_{t=1}^T c_{smt}} \quad (2)$$

where  $c_{smt}$  is the probability of being in state  $s$  with the mixture component  $m$  at time  $t$  given that the HMM with  $\Lambda$  generates the observation vector  $x_t$ ,  $\nu_{sm}$  is calculated using  $w_{sm}$  and  $\tau_{sm}$  according to eq. (3), and  $\tau_{sm}$  is set experimentally.

$$\nu_{sm} = w_{sm} \sum_{m=1}^M \tau_{sm}. \quad (3)$$

In MCE estimation [5][6], the discriminant function  $g_k(\cdot)$  and the misclassification function  $d_k(\cdot)$  for class  $k$  are defined by eqs. (4) and (5).

$$g_k(X, \Lambda) = \log[\mathcal{L}_k(X)], \quad (4)$$

$$d_k(X, \Lambda) = -g_k(X, \Lambda) + G_k(X, \Lambda) \quad \text{and} \quad (5)$$

$$G_k(X, \Lambda) = \log\left\{\frac{1}{K-1} \sum_{j,j \neq k} \exp[\eta g_j(X, \Lambda)]\right\}^{\frac{1}{\eta}},$$

where  $\mathcal{L}_k(\cdot)$  is the likelihood function for class  $k$ . Then,  $\Lambda$  is estimated by the segmental GPD (segmental generalized probabilistic descent) algorithm [7] so as to minimize the misclassification function  $d_k(\cdot)$ . In practice, the class loss function  $l(d_k)$  in eq. (6), which can be differentiated, is used instead of  $d_k(\cdot)$ , and  $\Lambda$  is updated iteratively by eq. (7).

$$l(d_k) = \frac{1}{1 + \exp(-\phi(d_k + \psi))} \quad \text{and} \quad (6)$$

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t \nabla l(d_k), \quad (7)$$

where  $\epsilon_t$  is adjusted for each iteration in order that  $\Lambda$  falls into one of the local minima near the initial model parameters. For each iteration,  $\epsilon_t$  is scaled with a proper range, the phoneme recognition experiment using the training data is performed for each scaled  $\epsilon_t$ , and the final  $\epsilon_t$  of the iteration is set to the  $\epsilon_t$  that achieves the highest performance (Figure 1).

### 3 EXPERIMENTAL CONDITIONS

In the experiments, 256-tied-mixture HMMs were used as context-independent phoneme HMMs. The number of phoneme HMMs was 43. Sentence data uttered by five male speakers in the ATR database of set B and sentence data uttered by 30 male speakers in the Acoustical Society of Japan (ASJ) database - in total 7,016 sentences - were used for estimating the parameters of speaker-independent phoneme HMMs by the Baum-Welch algorithm.

For speaker adaptation and evaluation, the sentence data uttered by six male speakers in the ATR database of set C were used and the average phoneme recognition rates over the speakers were used for evaluation. In the phoneme recognition experiments, each utterance was segmented by the Viterbi alignment according to the correct phoneme sequence in the sentence, the boundaries of the segments were assumed to be correct, and then the phoneme HMM that had the maximum likelihood value for each segment was used as the recognition result of the segment. For speaker adaptation, either 1, 5, 10 or 50 sentences were used for each speaker. The experiments were performed for two sets of sentence data; the same data as for speaker adaptation ("closed") and a different set of data of 100 sentences ("open"). The cepstral coefficients were calculated by LPC analysis with an order of 16, a frame period of 8 ms, and a frame length of 32 ms.

The number of training iterations was 10 for MCE estimation. The mixture mean and weighting factors of phoneme HMMs were estimated for MAP and MCE estimation methods. In MAP estimation, the parameters were updated after processing all sentences for each iteration. In MCE training, the parameters were updated at the end of processing each

Table 1. Phoneme recognition rates for each speaker using open data. (\*: top score in three methods)

| speaker        | SI   | training data |       |         |             |       |         |             |       |         |
|----------------|------|---------------|-------|---------|-------------|-------|---------|-------------|-------|---------|
|                |      | 1 sentence    |       |         | 5 sentences |       |         | 10 sentence |       |         |
|                |      | MAP           | MCE   | MAP+MCE | MAP         | MCE   | MAP+MCE | MAP         | MCE   | MAP+MCE |
| <i>S1</i>      | 50.2 | 54.2          | 53.5  | *58.5   | 57.1        | 57.4  | *59.6   | 62.0        | 60.5  | *62.2   |
| <i>S2</i>      | 36.6 | 37.0          | 37.7  | *45.2   | 42.2        | *47.2 | 46.1    | 52.5        | 49.3  | *55.5   |
| <i>S3</i>      | 45.1 | 49.9          | 50.7  | *53.4   | 54.3        | 56.0  | *58.3   | 59.9        | 60.0  | *61.2   |
| <i>S4</i>      | 32.1 | 44.2          | 48.0  | *49.4   | 51.0        | 50.4  | *55.9   | 54.0        | 59.4  | *59.7   |
| <i>S5</i>      | 49.4 | 50.9          | *54.2 | 53.9    | 56.7        | *60.1 | 59.5    | 63.1        | *64.8 | 64.1    |
| <i>S6</i>      | 43.2 | 46.4          | 49.1  | *52.3   | 52.9        | 55.2  | *55.8   | 58.3        | *59.9 | 59.6    |
| <i>Average</i> | 42.8 | 47.1          | 48.9  | 52.1    | 52.4        | 54.4  | 55.9    | 58.3        | 59.0  | 60.4    |

sentence. For MAP estimation, the initial values of

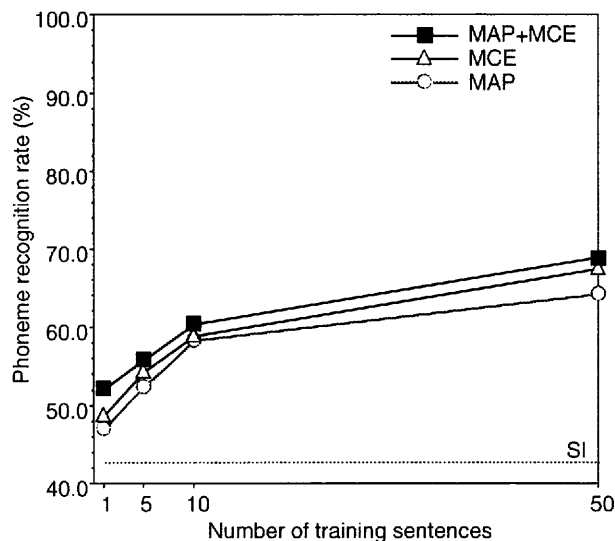


Figure 2. Phoneme recognition rates for open data.

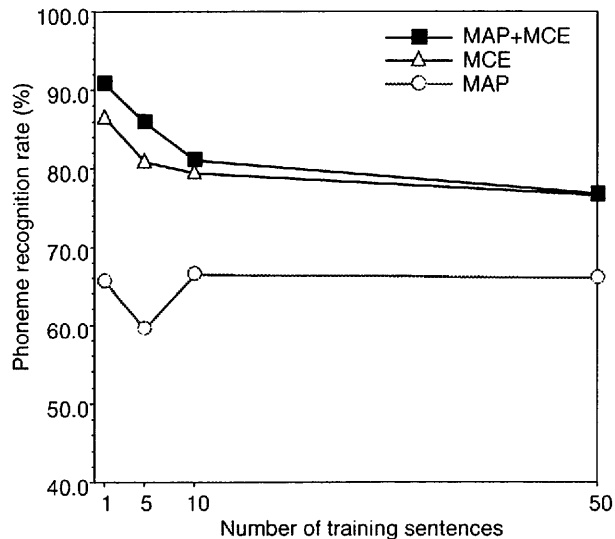


Figure 3. Phoneme recognition rates for closed data.

$\tau_{sm}$  were set to 1.5 when using 1 sentence for speaker adaptation, 1.1 when using 5 sentences, 0.9 when using 10 sentences, and 0.5 when using 50 sentences.

## 4 RESULTS

Figures 2 and 3 compare the average phoneme recognition rates over speakers directly using a speaker-independent phoneme HMM ("SI"), using only MAP estimation ("MAP"), using only MCE training ("MCE"), and using the combination method ("MAP+MCE").

For all numbers of training sentences, "MAP+MCE" performed best, "MCE" less well, and "MAP" least well for both open and closed data. Especially for closed data, the difference in performance between the methods of applying MCE estimation and "MAP" was large. These results indicate that MCE estimation is effective. Since MCE estimation aims at directly minimizing the recognition error, it can be considered that "MCE" performed better than "MAP."

Table 1 lists the phoneme recognition rates for each speaker using open data. The order for performance of the three methods for each speaker was almost the same as that for the average over all speakers.

Although over-tuning to the small amount of training data in MCE estimation is generally supposed to be a serious problem, even when only one sentence was used for speaker adaptation, "MCE" and "MAP+MCE" performed well for open data. For MCE estimation, since the model parameters were estimated to fall into one of the local minima near the speaker-independent model parameters or the model parameters adapted by MAP estimation, this problem was alleviated to some extent. These results indicate that MCE estimation even having speaker-independent models as initial conditions works very well for a small amount of training data.

## 5 DISCUSSION

To investigate the effectiveness of applying MCE estimation, experiments using the combination of ML and MCE estimation were conducted. In ML estimation, the parameters of speaker-independent phoneme HMMs were adapted to the speaker by using the Baum-Welch algorithm.

Figure 4 compares the average phoneme recognition rates using only ML estimation ("ML"), using the combination of ML and MCE estimation ("ML+MCE"), "MAP," and "MAP+MCE" for open and closed data when using 1 or 10 sentences for speaker adaptation. Although "ML" and "ML+MCE" performed well for closed data, for open data, they were much worse than "MAP" and "MAP+MCE." These results can be attributed to over-tuning to the small amount of training data and indicate that the effectiveness of MCE estimation is dependent on that of the initial models. When reasonable initial models such as those estimated by MAP can be used, MCE estimation can be considered to be an effective fine-tuning method of HMM parameters.

## 6 CONCLUSIONS

This paper investigated speaker adaptation of speaker-independent phoneme HMMs based on MCE training. A comparison of the performances of MAP, MCE, and the combination of MAP and MCE showed that the combination is the most effective of these three methods.

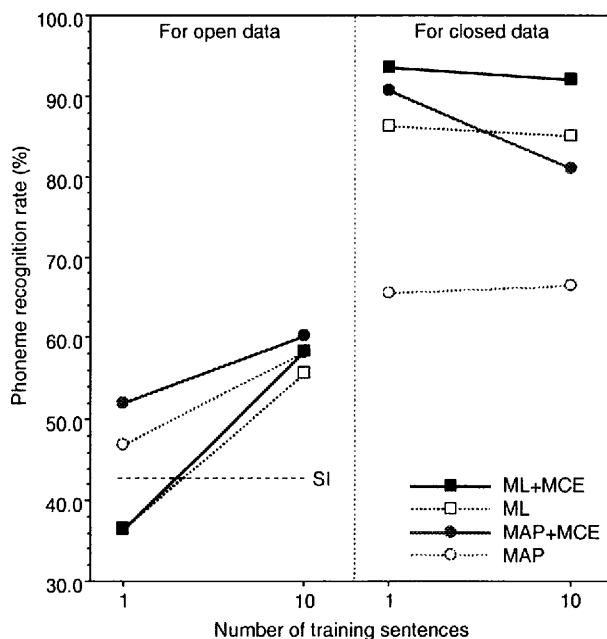


Figure 4. "ML+MCE" vs. "MAP+MCE."

Further study includes an improved method for finding better local minima that are robust for open data by using some constraints for MCE estimation. Moreover the performance of the combination method will be confirmed when using a huge number of context-dependent phoneme models and a parameter-tying technique for speaker adaptation. Although we got similar results over six speakers in the experiments, our method will be examined using a larger number of speakers.

## 7 ACKNOWLEDGMENT

The authors wish to acknowledge the members of the Furui Research Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions.

## REFERENCES

- [1] C.-H. Lee and J.L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," *Proc. ICASSP, Minneapolis*, pp. II-558-561, 1993.
- [2] J.L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp.291-298, 1994.
- [3] C.H. Lin, P.-C. Chang, and C.-H. Wu, "An initial study on speaker adaptation for Mandarin syllable recognition with minimum error discriminative training," *Proc. ICSLP, Yokohama*, pp. I-307-310, 1994.
- [4] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on ASSP, April*, 1991.
- [5] C.-S. Liu, C.-H. Lee, and W. Chou, "A study on minimum error discriminative training for speaker recognition," *J. Acoust. Soc. Amer.*, 97, 1, pp. 637-648, 1995.
- [6] B.-H. Juang and S. Katagiri, "Discriminative training," *J. Acoust. Soc. Jpn.*, 13, 6, pp. 333-339, 1992.
- [7] W. Chou, B.-H. Juang, and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," *Proc. ICASSP, San Francisco*, pp. I-473-476, 1992.