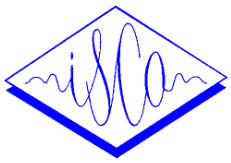


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

AN UNSUPERVISED SPEAKER ADAPTATION METHOD FOR CONTINUOUS PARAMETER HMM BY MAXIMUM A POSTERIORI PROBABILITY ESTIMATION

Yutaka Tsurumi and Seiichi Nakagawa

Toyohashi University of Technology,
Department of Information and Computer Sciences
Tempaku-cho, Toyohashi, 441, Japan

ABSTRACT

We studied an unsupervised speaker adaptation method on the sequential training that used the theory of *Maximum A Posteriori probability* (MAP) estimation for continuous parameter hidden Markov model (HMM). In this method, we should only specify the syllable label sequence for the utterance. The label sequences were provided automatically by the recognizer using a speaker independent (or adapted) model in advance and the language model. The syllable recognition rate using a language model for a given task is expected to have an accuracy of more than 90%, even if we use a *speaker independent* (SI) model. The experimental results on continuous speech recognition, for sentence or syllable, showed that the better initial model gave a performance comparable to that of supervised adaptation.

1 INTRODUCTION

It is hard to collect sufficient speech data for training a *speaker dependent* (SD) model from the same speaker. In contrast, to train a SI model we need not collect a large amount of speech data per speaker but from many speakers. A speaker adaptation technique is to adapt the SI model to a new speaker's model by the use of a small amount of training data[1].

For a *supervised speaker adaptation* (SSA) of continuous parameter HMM, the sequential concatenation training has been developed as a method of training that avoids hand-labelling[2]. The training is performed by using concatenated HMMs corresponding to the utterance' syllable (phoneme) label sequence.

Before, we have studied on such a training method using Viterbi segmentation/labeling to estimate the most suitable label sequence for a given utterance, and reported its details in [3]. In that study, we applied the theory of MAP estimation to the mean vector' and covariance matrix' adaptation. Using MAP estimation, even if there is only one sample, we can obtain the best parameters. Further, it is possible to again perform successive training for the refined HMM or the conventional ones. Therefore, this is more realistic and effective for on-line continuous speech recognition systems.

In this paper,we extend the adaptation algorithm to an *unsupervised speaker adaptation* (USA). The point here is that, unlike the previous method SSA, the label sequences are provided automatically by the recognizer which uses a standard SI model in advance. Thus, the necessary materials to train a USA model, in short, are only the utterances of the speakers to be adapted, which are untranscribed . This is based on the fact that the syllable recognition rate might reach an accuracy of about 90%, even if the sentence recognition rate in a speaker independent model is just about 40%.

The HMMs used in this study are the Japanese syllable-based ones (113 categories), and each HMM has a full covariance Gaussian probability density function for each state. We prepared three standard SI models in our experiments, and compared the effects of SSA and USA. One was the most basic set of HMM trained by *Maximum Likelihood* (ML) estimation on hand-labelled sentence utterances from 6 male speakers (named model-SI). Another was given by a successive training using MAP estimation from 30 male speakers (model-SI30), and a third one had additional dynamic features as acoustic feature parameters (only melcepstrum coefficients for model-SI and model-SI30). It was named model-SI30_RGC.

Recognition performances were evaluated on sentence recognition and continuous syllable recognition using $O(n)$DP [4]. Here, 84 sentences on "Sight-seeing guidance for Mt. Fuji" that were spoken by 6 male speakers were used for the tests. The other 20 sentences per speaker from the same task were used for the speaker adaptation.

Matsuoka and Lee have, as reported in [5], used almost the same USA method for context dependent phone-like HMMs, numbering more than 2000 categories with each having mixture Gaussian state observation distributions. However, their investigation showed no improvement as compared to SSA because there were too many models and all were not trained.

2 PARAMETER ESTIMATION

2.1 Concatenation Training [3]

The traditional concatenation training is as follows: at first, a sentence-based HMM is composed of concatenated syllable-based HMMs corresponding to the syllable label sequence for the sentence. Second, Viterbi segmentation is performed by using the Viterbi algorithm between the sentence utterance and the sentence-based HMM, and the frame samples corresponding to a state are spotted. And then parameters are reestimated by MAP estimation using the frame samples. Finally, the

sentence-based HMM is separated into each syllable-based HMM after the adaptation.

2.2 MAP Estimation

MAP estimation[6] is also called *Bayesian Succesive Estimation*[7]. This sequential training method is supervised. Therefore, we can estimate the parameter vector $\boldsymbol{\Theta}$ to have a maximum posterior probability for a given sentence. Given samples $\boldsymbol{X} = \{ \boldsymbol{X}_1 \cdots \boldsymbol{X}_N \}$ with each of them observed successively one by one, a posteriori probability is given by

$$P(\boldsymbol{\Theta}|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N) = \frac{P(\boldsymbol{X}_N|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{N-1},\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{N-1})}{\int P(\boldsymbol{X}_N|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{N-1},\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{N-1})d\boldsymbol{\Theta}} \quad (1)$$

We explain the estimation approach to calculate a mean vector and covariance matrix of a Gaussian distribution in the following:

(a) Estimation of a mean vector

For eq.(1), let us choose $\boldsymbol{\mu}$ as $\boldsymbol{\Theta}$, where $\boldsymbol{\mu}$ is the mean vector and is the object to be calculated in here. The density function $P(\boldsymbol{X}_1|\boldsymbol{\mu})$ is assumed to be normal with the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Note here that $\boldsymbol{\Sigma}$ is known (i.e. the covariance matrix of standard model-SI).

$$P(\boldsymbol{X}_1|\boldsymbol{\mu}) \simeq N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (2)$$

We assumed that the priori density function of the mean vector $\boldsymbol{\mu}$ is normal with the expected vector $\boldsymbol{\mu}_0$ and covariance matrix K_0, as follows:

$$P(\boldsymbol{\mu}) \simeq N(\boldsymbol{\mu}_0, \mathrm{K}_0) \quad (3)$$

Then substituting the above probabilities into eq.(1), after observing the first sample $\boldsymbol{X}_1$,

$$\begin{aligned} P(\boldsymbol{\mu}|\boldsymbol{X}_1) &= \frac{P(\boldsymbol{X}_1|\boldsymbol{\mu})P(\boldsymbol{\mu})}{\int P(\boldsymbol{X}_1|\boldsymbol{\mu})P(\boldsymbol{\mu})d\boldsymbol{\mu}} \simeq N(\boldsymbol{\mu}_1, \mathrm{K}_1) \\ &= C\exp[-\frac{1}{2}(\boldsymbol{X}_1-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}) \\ &- \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\mathrm{K}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)] \end{aligned} \quad (4)$$

, where eq.(4) indicates the term to be related to $\boldsymbol{\mu}$, thus C is constant. The estimated mean vector $\hat{\boldsymbol{\mu}}_1$ and covariance matrix $\hat{\mathrm{K}}_1$ are as follows:

$$\begin{cases} \hat{\boldsymbol{\mu}}_1 &= \mathrm{K}_0(\mathrm{K}_0+\boldsymbol{\Sigma})^{-1}\boldsymbol{X}_1 + \boldsymbol{\Sigma}(\mathrm{K}_0+\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}_0 \\ \hat{\mathrm{K}}_1 &= \mathrm{K}_0(\mathrm{K}_0+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma} \end{cases} \quad (5)$$

K_0 is a covariance matrix to indicate the extent of $\boldsymbol{\mu}$'s uncertainty before the estimation. In [6], it was calculated from the mean vector of each mixture of the standard model before the adaptation. But it is impossible to apply such a method in case of a single mixture distribution. So, we introduce an adaptation parameter α instead of K_0 and determine the value experimentally[3].

$$\mathrm{K}_0 = \alpha^{-1}\boldsymbol{\Sigma} \quad (6)$$

Transform the formula (5) using this α,

$$\hat{\boldsymbol{\mu}}_1 = \frac{\alpha\boldsymbol{\mu}_o + \boldsymbol{X}_1}{\alpha+1} \quad (7)$$

The estimated value after given N samples is shown as

$$\begin{aligned} \hat{\boldsymbol{\mu}}_N &= \frac{(\alpha+N-1)\boldsymbol{\mu}_{N-1} + \boldsymbol{X}_N}{\alpha+N} \\ &= \frac{\alpha\boldsymbol{\mu}_o + \sum_{i=1}^{N}\boldsymbol{X}_i}{\alpha+N} \end{aligned} \quad (8)$$

We kept α fixed through all states of all over syllable categories. In this paper, we chose the value $\alpha = 15$ by following literature[3].

(b) Estimation of covariance matrix

In this case, we should estimate two parameters (that is, a mean vector and a covariance matrix), so a priori distribution and a posteriori probability form a joint probability function. The detailed explanation is in [7]. The estimated values by using one or N samples are

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_1 &= \frac{\boldsymbol{X}_i\boldsymbol{X}_i^T - (\alpha+1)\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T + \beta\boldsymbol{K}_0 + \alpha\boldsymbol{\mu}_o\boldsymbol{\mu}_o^T}{\beta+1} \\ \hat{\boldsymbol{\Sigma}}_N &= \frac{1}{\beta+N}\Big\{\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T - (\alpha+N)\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T \\ &+ \beta\boldsymbol{K}_0 + \alpha\boldsymbol{\mu}_o\boldsymbol{\mu}_o^T\Big\} \\ &= \frac{1}{\beta+N}\Big\{\boldsymbol{X}_N\boldsymbol{X}_N^T - (\alpha+N)\boldsymbol{\mu}_N\boldsymbol{\mu}_N^T \\ &+ (\beta+N-1)\boldsymbol{\Sigma}_{N-1} \\ &+ (\alpha+N-1)\boldsymbol{\mu}_{N-1}\boldsymbol{\mu}_{N-1}^T\Big\} \end{aligned} \quad (9)$$

,where β is a coefficient. According to [7], β is based on a number of samples to estimate a covariance matrix of standard HMM. But in this paper, β was determined experimentally as α and is set at 50. The formula for a mean vector is the same as (8).

3 UNSUPERVISED ADAPTATION

Speaker adaptation is performed with successive training of a SI model using a small amount of adaptation speech. There are two adaptation methods. One, well known as the supervised speaker adaptation (SSA), achieves the adaptation in accordance with correct label sequences. Another method is known as the unsupervised speaker adaptation (USA). In this method, the

label sequences were provided automatically by the sentence recognizer using a SI model and language model. Therefore, our adaptation system consists of a sentence recognition part and an adaptation part as shown in Figure 1.

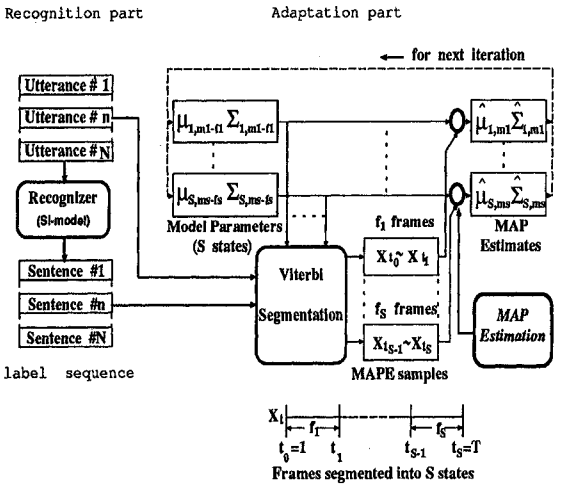


Figure 1. Block diagram of Unsupervised Speaker Adaptation

All of the utterances of an adapted speaker is recognized by using a SI model in advance. The reason why we chose this procedure is that we want to investigate the relationship between an average of an initial sentence recognition rate by SI model and the recognition performance of after adaptation. For practical use, recognition and adaptation will be done alternately[5]. Such scheme might yield better results. Whenever the adaptation data (i.e result of sentence recognition) corresponding to the adaptation speech is obtained, it estimates the best parameters sequentially.

4 EXPERIMENTS

4.1 Speech Analysis and Material

The speech corpora were sampled at 12KHz. We used 14 LPC cepstrum coefficients and signal power for 8ms. 10 LPC mel-cepstrum coefficients were used as the feature parameter.

There were three standard SI models used here as described in Section 1 and are compared for the effects of SSA and USA. Model-SI was the most basic set of HMM trained by ML estimation on hand-labelled sentence utterances from 6 male speakers. These were collected from ATR Japanese continuous speech database, A-J sets, and from 216 words in the same database for certain categories which do not have enough samples. The number of samples per category was limited to 1200 samples. Model-SI30 was given a successive training by MAP estimation from 30 male speakers, and model-SI30_RGC had additional dynamic features. The HMMs used in these experiments were syllable-based continuous full covariance Gaussian distribution HMMs with discrete distributed duration control. Every model had 5-states, and it was represented as a left-to-right transition structure.

For the acquisition of a syllable label sequence, we applied the continuous speech recognizer using a top-down parser that was activated by the One-Pass search algorithm, Earley-like parser for Context Free Grammar and a dynamically changing finite state network[8].

Recognition performances were evaluated on sentence recognition and $O(n)$DP continuous syllable recognition. Here, 84 sentences on "Sight-seeing guidance for Mt. Fuji" that were spoken by 6 male speakers who belong to our laboratory (AK, MM, SA, TK, TS, YM) were used for the tests. The remaining 20 sentences per speaker from the same task were used for the speaker adaptation.

4.2 Experimental Results

We prepared three sets of initial HMM: model-SI, SI30 and SI30_RGC. In this experiment, we compared the results using the SI models with those results using USA or SSA model.

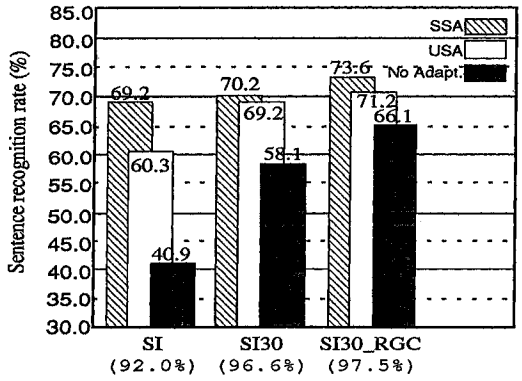
We recognized 20 sentences per speaker for an unsupervised training at the beginning. Table 1 shows the recognition rates of sentence and the syllable label sequences that were given from a recognized sentence. If they are 100% correct, it will be the same method as SSA. Table 1 shows that the syllable recognition rate reaches an accuracy of about 90–96%, even if the sentence recognition rate in a speaker independent model is about 40–60%. On the other hand, there were few insertions or deletions, so average segmentation rates were 96–98%. We can know from the table that the best quality speaker independent model is SI30_RGC.

After the USA training for these initial models have been performed, we investigated the recognition performance. Figures 2(a) and (b) illustrate the sentence recognition rate and the syllable recognition rate using O(n)DP, respectively. The results for SI model (no adaptation) and SSA are also shown for comparison. Note that the values in parentheses under the model type denote the average of syllable recognition rates in label sequences.

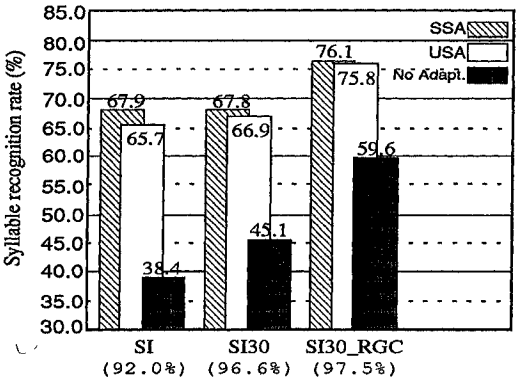
When the initial model SI was used, the difference between the USA and SSA was 8.9%. But using the model SI30, the difference was kept at only 1.0%. And the evaluation by means of syllable recognition rate shows a good result very similar to the SSA case. Using SI30_RGC as the initial SI model only gave a 0.3% dif-

Table 1. Label sequence recognition for adaptation (20 sentences, 6 males, unit:%)

Recognition unit	Used model		
	SI	SI30	SI30_RGC
sentence	48.3	62.5	66.7
syllable	92.0	96.6	97.5



(a) Effectiveness by sentence recognition rate



(b) Effectiveness by continuous syllable recognition rate

Figure 2. Effectiveness of Unsupervised Adaptation (values in parentheses denote average syllable recognition rates of adapting utterances)

ference. These remarkable results were caused by that the syllable label sequences for the adaption had high reliability of 96.6% or 97.5%. Thus, we could say that the method described above is very effective. According to our other report [9], if the syllable label sequence accuracy reaches about 90% (after the conversion into sentence recognition, it is about 40%), we can take advantage of this unsupervised adaptation. When the task dependent sentences were used for the adaptation, the recognition rate of after adaptation converges when about 20~30 sentences were given as our experiments.

We also performed an experiment which used context dependent HMM(208 models)[3] instead of these SI HMMs(113 models). We have prepared the syllable label sequence that the recognition rate was 92% for the adaptation. Then, the syllable recognition rates of no-adaptation, after USA and SSA approaches were 40.2%, 67.5% and 69.0% respectively. So there were effects of USA approach in this case with many model parameters too (see SI in Figure 2(b)).

5 SUMMARY

Based on the theory of MAP estimation, we can adapt syllable HMMs by using only the syllable label sequence for the adaptation. In this study, we provide the label sequences automatically obtained by the recognizer which uses a standard speaker independent (SI) model in advance. We compared the effects of unsupervised speaker adaptation (USA) approaches with supervised speaker adaptation (SSA) using three initial SI models. The USA results on continuous speech recognition showed that the better initial model gave a performance comparable to that of the SSA case. Therefore, it is more realistic and effective for on-line continuous speech recognition systems.

In further research, only more reliable recognized results should be applied to spontaneous speech.

References

[1] Y.Hirata, S.Nakagawa, "Speaker adaptation of continuous parameter HMM", Proc. ICSLP, pp.377–380, 1990.

[2] Maruyama, T.Hanazawa, T.Kawabata, K.Shikano, "English Word Recognition Using HMM Phone Concatenated Training", Tech. Report of IEICE, SP88-119, 1989 (in Japanese).

[3] S.Nakagawa, T.Koshikawa, "Speaker Adaptation for Continuous Parameter HMM by maximum A Posteriori Probability Estimation", Jour. ASJ, Vol.49, No.10, pp.721–728, 1993 (in Japanese).

[4] S.Nakagawa, "A connected spoken word recognition method by $O(n)$ dynamic programming pattern matching algorithm", Proc. ICASSP, pp.296–299, 1983.

[5] T.Matsuoka, Chin-Hui Lee, "A Study of On-line Bayesian Adaptation for HMM-Based Speech Recognition", Proc. Eurospeech, pp.815–818, 1993.

[6] Chin-Hui Lee et al., "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models",IEEE Trans. Signal Processing, Vol.39, pp.806–814, 1991.

[7] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Edition, Academic Press, 1990.

[8] S.Nakagawa, A.Kai, "A Frame-Synchronous Continuous Speech Recognition Algorithm Using A Top-Down Parsing of Context-Free Grammar",Proc. ICSLP, pp.257–260, 1992.

[9] Y.Tsurumi, S.Nakagawa, "Unsupervised Speaker Adaptation for Continuous Parameter HMM by Maximum A Posteriori Probability Estimation", Tech. Report of IEICE, SP93-104, 1993 (in Japanese).