Workshop on Multimedia Technologies

Speaker Adaptation Based on MAP Estimation Using Fuzzy Controller

Yau-Tarng Juang*, Kuo-Chang Huang, and Ing-Jr Ding
Department of Electrical Engineering, National Central University,
Chung-Li, Taiwan 32054, R.O.C.

Abstract

This paper presents a technical incremental speaker adaptation method called FCMAP, which incorporated an appropriate fuzzy controller into maximum a *posteriori* (MAP) to adapt the hidden Markov model parameters. The recognition performance of MAP is consistently improved and stabilized by an appropriate fuzzy controller. The recognition results obtained in unsupervised adaptation experiments showed that FCMAP estimation was effective even when only few utterances from a new speaker were used for adaptation.

Keywords: Hidden Markov model; Maximum a *posteriori* estimation; Maximum likelihood

---

The work was partially supported by the National Science Councils of the Republic of China under the contract NSC89-2218-E-008-004.

*Corresponding author. Fax: 886 3 4255830; E-mail address: ytjuang@ee.ncu.edu.tw (Yau-Tarng Juang).

1. Introduction

Many techniques compensating the degradation caused by mismatches between the training and test condition have been developed. They are roughly grouped into two categories, namely 1) feature compensation (Lee, 1998), in which the process of feature extraction is modified and 2) model adaptation (Gauvain and Lee, 1994; Leggetter and Woodland, 1995), in which the parameters of recognition models are adjusted. Although combining these two techniques has been shown effective (Sankar and Lee, 1996), we focus our discussion in the present study on model adaptation. Model adaptation techniques are an efficient way to reduce the mismatch that typically occurs between the training and test condition of any speech recognizer. Adaptation techniques can usually be divided into two families of approaches. On one hand, direct model adaptation attempts to directly reestimate the model parameters, for example using Maximum a *posteriori* (MAP) adaptation (Gauvain and Lee, 1994; Zavagliogkos et al., 1995). Since direct adaptation only reestimates model parameters of the corresponding units appearing in the adaptation data, a large amount of such data is needed to observe any significant improvement in performance. However, nice asymptotic properties are usually observed, meaning that the performance improves as the amount of adaptation data increases. On the other hand, indirect model adaptation (Maximum Likelihood linear Regress, MLLR) (Leggetter and Woodland, 1995) applies a general transformation on some clusters of model parameters.

Because each individual model is transformed, the approach is quite effective when a small amount of adaptation data is available. However, as the amount of adaptation data increases, the performance improvement quickly saturates.

MAP estimate has been successfully applied to speaker adaptation in speech recognition system using hidden Markov models. When the amount of data is sufficiently large, MAP estimation yields recognition performance as good as that obtained using maximum-likelihood estimation. This paper describes a fuzzy controller application to maximum a *posteriori* approach to improve the MAP estimates obtained when the amount of adaptation data is small. The adaptation scheme is described in more detail in the following section, and is shown to be capable if giving reasonable improvements in performance with a very little amount of adaptation data. In section 2, the theoretical formulation of the MAP estimation is briefly described. Next, we introduce a fuzzy controller algorithm for adjusting the parameter of the MAP estimation (FCMAP) in section 3. The effectiveness of the FCMAP algorithm was demonstrated in a set of unsupervised adaptation experiments. We report on experimental results with different adaptation scenarios in section 4. Finally, we summarize our findings in section 5.

2. Overview of MAP estimation

A Bayesian adaptation training procedure has been applied with good success for model-based recognizers using continuous density hidden Markov model (CDHMM). Speaker adaptation of the CDHMM parameters can usually be formulated as a Bayesian learning procedure. Besides, a Bayesian learning procedure (Lee et al., 1991) is easily integrated into the segmental *k*-means training procedure (Juang, 1990) for obtaining adaptive estimates of the CDHMM parameters (Rabiner et al., 1986; Lee et al., 1991).

The segmental *k*-means algorithm with embedded Bayesian adaptation consists of the following two steps:

1) Obtain the optimal state segmentation of a given observation sequence *Y*, based on a given model $\hat{\lambda}$, i.e.,

$$\hat{s} = \arg\max_{s} P(Y, s \mid \hat{\lambda}) P_0(\hat{\lambda}) \tag{1}$$

where $P_0(\hat{\lambda})$ is the prior distribution of the parameter $\hat{\lambda}$

2) Based on the optimal state sequence $\hat{s}$, find the MAP estimate

$$\hat{\lambda} = \arg\max_{\lambda} P(Y, \hat{s} \mid \lambda) P_0(\lambda). \tag{2}$$

These two steps are iterated until some fixed-point solution is reached.

As usual, when adapting parameters of CDHMMs, it is more effective and simplificative to adapt the mean vectors of Gaussian distributions only than other parameters such as the variance, transition probability, and so on. Thus, in this paper,

only Bayesian adaptation of the mean of Gaussian mixture is utilized. Let μ and σ^2 be the mean and the variance parameters of one component of a state observation distribution. Assume the mean μ is random with a *priori* distribution $P_0(\mu)$, and the variance σ^2 is known and fixed. The conjugate *priori* for μ is also Gaussian with mean ν and variance $\tilde{\tau}^2$. If the conjugate *priori* for the mean is utilized to perform the Bayesian adaptation, then the MAP estimate for the parameter is solved by (Duda and Hart, 1973)

$$\hat{\mu}_P = \frac{n\tilde{\tau}^2}{\sigma^2 + n\tilde{\tau}^2} m_p + \frac{\sigma^2}{\sigma^2 + n\tilde{\tau}^2} \nu \tag{3}$$

which could be alternatively expressed as follows by (Tonomura et al., 1995)

$$\hat{\mu}_p = \frac{n}{\tau + n} m_p + \frac{\tau}{\tau + n} \mu_p^I, \ \tau = \frac{\sigma^2}{\tilde{\tau}^2} \tag{4}$$

where m_p is the sample mean of the p-th Gaussian distribution, μ_p^I is the p-th mean vector of the Gaussian distribution of the initial model, $\hat{\mu}_p$ is the p-th mean vector of the Gaussian distribution of the adapted model, n is the total number of training samples observed for the corresponding Gaussian mixture component , and τ is the relative balance between the a *priori* knowledge and empirical data.

3. A fuzzy controller application (FCMAP)

In Eq. (4), the parameter τ is regarded as the control parameter for the adaptation speed (Takahashi and Sagayama, 1995). When τ is large, the *priori*

density is sharply peaked around the values of the seed HMM parameters which will be only slightly modified by the adaptation process. Conversely, if τ is small the adaptation will be very fast. The parameter τ is a *priori* density parameter that controls the balance between prior information and the new training data. Eq. (4) shows in such a way that the estimated mean vector is obtained as the interpolated one using the initial mean weight by τ and the sample mean weighted by the effective sample counts *n*. In the literature (e.g. Lee and Gauvain, 1993), they used a common value τ for all the Gaussians of a given state, or for all states of an HMM, or even for all HMMs. In our motivation, to further increase the robustness, the τ value can be modified according to the amount of adaptation data. Therefore, a fuzzy controller was utilized to determine the appropriate τ, and it is described as follows. The input of the fuzzy controller is N, which is the total number of the training samples observed for all Gaussian mixture components.

Rule 1: If N is small,

Then τ is large.

Rule 2: If N is medium,

Then τ is medium.

Rule 3: If N is large,

Then τ is small.

Now, let three respective functions $f_1(N), f_2(N),$ and $f_3(N)$ be suitably designed functions corresponding to τ_L, τ_M, and τ_S, and $M_1(N)$, $M_2(N)$ and $M_3(N)$ be the corresponding membership functions of N for small, medium, and large training samples, respectively. The above-mentioned inference rules then become:

Rule 1: If N is $M_1(N)$,

Then $\tau_L = f_1(N)$.

Rule 2: If N is $M_2(N)$,

Then $\tau_M = f_2(N)$.

Rule 3: If N is $M_3(N)$,

Then $\tau_S = f_3(N)$.

Given an input N, the final output τ of the fuzzy system is inferred and defuzzified as follows

$$\tau = \frac{\sum_{i=1}^{3} M_i(N) f_i(N)}{\sum_{i=1}^{3} M_i(N)}. \tag{5}$$

4. Experiments and results

4.1 Database and System Description

This section summarizes the results of various experiments that were conducted to evaluate the proposed FCMAP algorithm. New adaptation and testing data from

five nonnative male speakers (labeled as A, B, C, D, and E) were recorded simultaneously under a close-talking microphone. The data for adaptation consisted of 30 utterances from each speaker. For testing, we collected from each speaker 60 utterances uttered twice for 30 city names. The speech signal was sampled 8kHz and the analysis frames were 30-ms wide with a 20-ms overlap. For each frame a 24 dimensional feature vector was extracted. The feature vector for each frame consisted of a 12-dimensional (12-D) mel-cepstral, a 12-D delta-mel-cepstral vector. Speaker-independent HMMs were trained using the MAT400 (Wang, 1997) training set consisting of 4800 utterances from native Mandarin talker (184 females and 216 males), each providing 12 utterances. For recognition, we used a set of 146 context-dependent sub-syllable HMMs. All units had three states (for initial part) or six states (for final part).

Experiments were focused on investigating how the performance improvement and adaptation speed are changed by the value of the parameter τ in word-by-word incremental speaker adaptation. The value of 30 for τ was used in the conventional MAP. The appropriate value of τ was chosen according to the amount of adaptation data in the proposed FCMAP approach.

The membership functions of the designed fuzzy controller are detail described as follows (Fig. 1):

$$M_1(N)=\begin{cases}1, & N \le N_1 \\ \dfrac{N_2-N}{N_2-N_1}, & N_1 \le N \le N_2 \\ 0, & N \ge N_2\end{cases}$$

$$M_2(N)=\begin{cases}0, & N \le N_1 \text{ or } N \ge N_3 \\ \dfrac{N-N_1}{N_2-N_1}, & N_1 < N \le N_2 \\ \dfrac{N_3-N}{N_3-N_2}, & N_2 \le N < N_3\end{cases}$$

$$M_3(N)=\begin{cases}0, & N \le N_2 \\ \dfrac{N-N_2}{N_3-N_2}, & N_2 < N < N_3 \\ 1, & N \ge N_3\end{cases}$$

And the three corresponding functions $f_i(N)'s$ are as follows：

$$f_1(N)=\frac{b}{\log(N)+a}$$

$$f_2(N)=\frac{N}{c}$$

$$f_3(N)=\frac{\log(N)}{N}$$

where *a*, *b* and *c* are some suitably chosen integers.

4.2 Results

Fig. 2 shows adaptive learning curves of the proposed FCMAP and conventional MAP. The average recognition rates are plotting along the adaptive word counts from 2 to 30. The performance of FCMAP and MAP (in the case of $\tau = 30$) is compared in Table 1 for isolated word recognition. The performance of FCMAP increased rapidly for all sizes of adaptation data and for all the speakers. Nevertheless, in conventional

MAP, the word recognition performance tended to degrade slowly to beneath the baseline of the initial models from about 2 words to 22 words and was nearly equal to the baseline form about 26 words. The FCMAP method showed better recognition accuracy than that obtained with the conventional MAP not only when the amount if data was large but also when the amount data was small. This is probably because parameter estimation was more robust than that in MAP, since a fuzzy controller was used. The results show that not only did an appropriate fuzzy controller accelerate the adaptation speed of the MAP but also improve and stabilize performance by the proposed FCMAP.

5. Conclusion

The FCMAP method for adaptation of hidden Markov model parameters enhances the performance of the conventional MAP method when the amount of data is small by utilizing an appropriate fuzzy controller. Its effectiveness was confirmed in a set of recognition experiments. The fuzzy controller improves and stabilizes the recognition performance of the conventional MAP. The proposed FCMAP method is simple to process and requires no pooling of the adaptation data.

References

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. New York : Wiley.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans. Speech Audio Processing. 2, 291–298.

Juang, B.-H., Rabiner, L.R., 1990. The segmental K-means algorithm for estimating parameters of hidden Markov models. IEEE Trans. Signal Processing 38(9), 1639-1641.

Lee, C.H., Lin, C.H., Juang, B.H., 1991. A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans. ASSP 39, 806–814.

Lee, C.-H., Gauvain, J.-L., 1993. Speaker adaptation based on MAP estimation of HMM Parameter. ICASSP93 II, 558-561.

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Commun. 25, 29–47.

Leggetter, C. J., Woodland, P. C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. 9, 171-185.

Rabiner, L.R., Wilpon, J.G., Juang, B.H., 1986. A segmental k-means training for connected word recognition. AT&T Tech. J. 65, 21-32.

Sankar, A., Lee, C.H., 1996. A maximum likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. SAP 4, 190-202.

Takahashi, J., Sagayama, S., 1995. Vector-field-smoothed Bayesian learning for incremental speaker adaptation. ICASSP-95 1, 696–699.

Tonomura, M., Kosaka, T., Matsunaga, S., 1995. Speaker adaptation based on transfer vector field smoothing using maximum a *posteriori* probability estimation. ICASSP-95 1, 688–691.

Wang, H.C., 1997. MAT – A project to collect Mandarin speech data through telephone networks in Taiwan. Computational Linguistics and Chinese Language Processing 2, 73-89.

Zavagliogkos, G., Schwartz, R., Makhoul, J., 1995. Batch, incremental and instantaneous adaptation techniques for speech recognition. In Proc. ICASSP, Detroit, MI, 676–679.

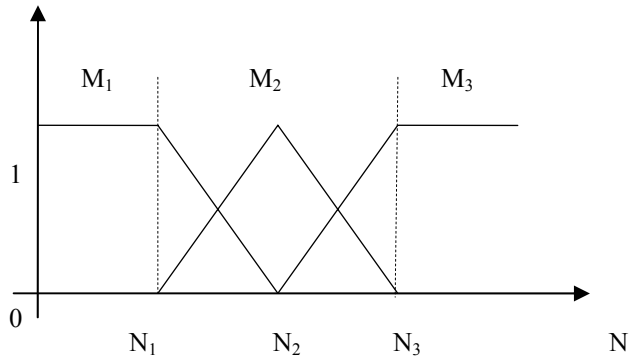

Fig. 1. The membership functions of fuzzy controller.

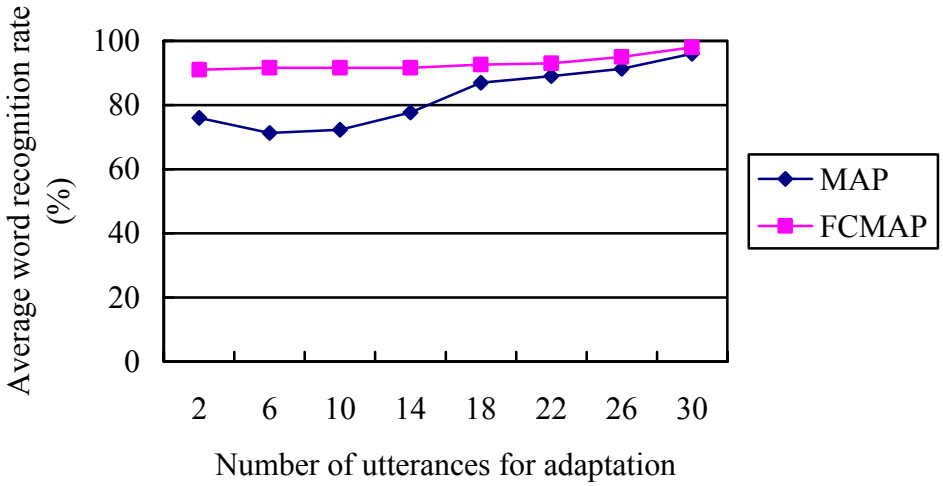

Fig. 2. Adaptive learning curves in word recognition.

Table 1
Recognition rates (%) of each speaker obtained by using conventional MAP (τ was set 30) and FCMAP.

Speaker	Adaptation Method	Recognition rate (%)								
		Number of utterance for adaptation								
		0	2	6	10	14	18	22	26	30
A	FCMAP	90	90	90	90	90	93.33	95	96.67	98.33
	MAP		53.33	38.33	48.33	68.33	81.67	86.67	90	93.33
B	FCMAP	90	90	90	90	90	90	91.67	95	100
	MAP		75	73.33	75	81.67	88.33	93.33	91.67	96.67
C	FCMAP	91.67	91.67	91.67	91.67	91.67	93.33	91.67	93.33	95
	MAP		83.33	75	70	71.67	83.33	88.33	90	96.67
D	FCMAP	93.33	93.33	95	95	95	93.33	95	96.67	100
	MAP		81.67	85	81.67	80	88.33	88.33	91.67	95
E	FCMAP	90	90	91.67	91.67	91.67	93.33	91.67	93.33	96.67
	MAP		86.67	85	86.67	86.67	93.33	88.33	93.33	98.33
Average	FCMAP	**91**	**91**	**91.67**	**91.67**	**91.67**	**92.66**	**93**	**95**	**98**
	MAP		**76**	**71.33**	**72.33**	**77.67**	**87**	**89**	**91.33**	**96**