NOVEL TRAINING METHOD FOR CLASSIFIERS USED IN SPEAKER ADAPTATION

Naoto Iwahashi

SONY D21 Laboratory
6-7-35 Kitashinagawa Shinagawa-ku, Tokyo, 141, Japan
naoto@av.crl.sony.co.jp

ABSTRACT

This paper describes a novel method for training a classifier that will perform well after it has been adapted to input speakers. For off-line (batch-mode) adaptation methods which are based on the transformation of classifier parameters, we propose a method for training classifiers. In this method, the classifier is trained while the adaptation to each speaker in the training data is being carried out. The objective function for the training is given based on the recognition performance obtained by the adapted classifier. The utility of the proposed training method is demonstrated by experiments in a five-class Japanese vowel pattern recognition task with speaker adaptation.

1. INTRODUCTION

In speech recognition, adequate recognition performance shuld be maintained in spite of changes in the input speaker. To achieve such robustness, several approaches by improving classifiers have been studied [1]. A conventional way to obtain robust classifiers is to use a large amount of training data from as many speakers as possible [2]. In this approach, although the average performance in speakers can be improved, the performance of individual speakers is not necessarily satisfactory. To find a way of achieving higher recognition performance, the adaptation of the recognizer to different speakers has been intensively studied [1]. Several methods have been proposed concerning adapted parameters, transformation methods, and optimization criteria for training the transformation. From among these methods, off-line (batch-mode) adaptation methods based on the transformation of classifier parameters have been investigated to persue efficient adaptation with a small amount of adaptation training data [3, 4]. In this transformation-based adaptation method, either speaker-dependent or independent classifiers have been used as the classifier that is adapted. However, these classifiers are not necessarily optimal with respect to recognition performance after adaptation is carried out on them. To improve the performance after adaptation, other types of classifiers have also been investigated [5]. However, due to the lack of a theoretical framework to formalize this situation, we have had to depend on a heuristic approach to seek better classifiers.

In this paper, we present a framework for optimizing classifiers so as to achieve high performance after adaptation has been carried out, and propose a training method based on this framework. In the proposed training method, the classifier is trained while adaptation is being carried out. The objective function for the training is given based on the recognition performance obtained by the adapted classifier. The proposed method makes any optimization criteria, such as minimum squared error (MSE), maximum likelihood (ML), minimum classification errors (MCE) or maximum mutual information (MMI), available for both classifier training and the adaptation processes. It is also applicable to adaptation to changes in other input conditions, such as background noise, room acoustics, or channel noise. In the following section, the problem will be formulated so as to handle changes in these input conditions as well as in the speaker.

2. PROBLEM FORMULATION

We consider the classification of a d-dimensional vector $\mathbf{O}$ in an observation space into one of the K classes $\{C_k\}_{k=1}^{K}$, with adaptation of the classifier to changes in the input conditions represented by the set of parameters V. This classification uses the decision rule:

$$C(\mathbf{O}) = C_i \quad \text{if} \quad g_i(\mathbf{O}; F(\Lambda, \Gamma)) = \max_j g_j(\mathbf{O}; F(\Lambda, \Gamma)), \quad (1)$$

where $C(\cdot)$ denotes a classification operation, g_i is the discriminant function for class C_i, Λ is the set of classifier parameters, Γ is the set of parameters of the adaptation transformation, and $F(\Lambda, \Gamma)$ represents the parameter set of the classifier obtained through the transformation of Λ by Γ. The block diagram of this pattern recognizer is shown in Figure 1. At the stage of adaptation to input condition V, the value of Γ is decided by using an adaptation training sample set extracted under V in an off-line manner. The ultimate goal here is to achieve discriminant functions that minimize the probabilities of classification error, given that an adaptation procedure is performed on these functions.

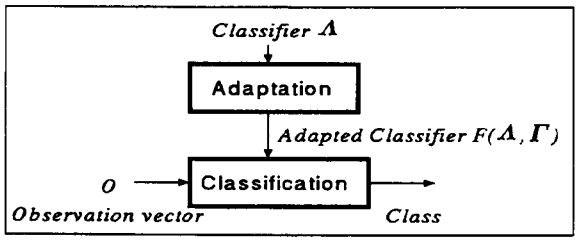

Figure 1: Pattern recognizer with adaptation of classifier

The problem of optimizing a conventional pattern classifier, which is designed independently of the adaptation procedure, can be formulated as follows:

$$\sum_{k=1}^{K} P(C_k) \int 1(g_k(\mathbf{O};\Lambda) \neq \max_i g_i(\mathbf{O};\Lambda)) p(\mathbf{O}|C_k) d\mathbf{O} \rightarrow \min, \quad (2)$$

where $P(C_k)$ is the class probability, $p(\mathbf{O}|C_k)$ is the conditional probability density function, and $1()$ is an indicator function:

$$1(\alpha) = \begin{cases} 1, & \text{if } \alpha \text{ is true} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The integration is over the entire observation space. Note that this formulation is for minimizing the expected error rate when the classifier has not been adapted to the input conditions.

Now, we consider the problem of optimizing the classifier used in rule (1). The classifier should be optimized so as to achieve high performance after adaptation has been carried out. This problem is formulated as follows:

$$\int p(V) \int \sum_{k=1}^{K} P(C_k|V) \times \int 1(g_k(\mathbf{O};F(\Lambda,\Gamma)) \neq \max_i g_i(\mathbf{O};F(\Lambda,\Gamma))) \times p(\mathbf{O}|C_k,V) p(\Gamma|\Lambda,V) d\mathbf{O} d\Gamma dV \rightarrow \min. \quad (4)$$

Note that because the value of Γ is decided in the adaptation process using extracted adaptation training samples, Γ can be taken as random variables. The conditional probability density $p(\Gamma|\Lambda,V)$ depends on the adaptation procedure used. Each integration is over the entire space of the corresponding variable.

3. TRAINING METHOD

We will now describe the method used to train a classifier, which is used in speaker adaptation, based on formula (4). In the case of speaker adaptation, we can assume that parameter V, which represents the input speaker, takes discrete values $(V_b, b = 1, 2, ...)$. The general scheme for the off-line adaptation method based on the transformation of classifier parameters is formulated first, and then the objective function for the training of a classifier is shown.

3.1. Formulation of Adaptation

In the process of adaptation to input speaker V_b, the values of parameters Γ_b of the adaptation transformation is determined using the adaptation training sample set (denoted by $\mathcal{A}_b$) extracted from utterances spoken by speaker V_b. The optimum value $\hat{\Gamma}_b$ of Γ_b is obtained as follows:

$$\hat{\Gamma}_b = \operatorname*{argmin}_{\Gamma} \sum_{k=1}^{K} \sum_{i=1}^{m_{b,k}} \ell_a(\mathbf{O}_{b,k,i}^{(\mathcal{A}_b)}, C_k, F(\Lambda,\Gamma)), \quad (5)$$

where $m_{b,k}$ is the number of samples in class C_k in $\mathcal{A}_b$, $\mathbf{O}_{b,k,i}^{(\mathcal{A}_b)}$ denotes the ith feature vector in them, and ℓ_a is the loss function for individual samples, which is used to determined the value of the transformation parameters for the adaptation. By changing the loss function, (5) can represent any optimization criteria, such as MSE, ML, MCE, MMI, and so on. Note that minimization in (5) is often carried out with some constraints among the parameters of the transformation, such as smoothing, to reduce small-sample-size effects[6].

3.2. Objective Function for Classifiers

Next, we consider the objective function to be minimized for training a classifier using the sample set $\mathcal{X}$ which is composed of samples extracted from utterances spoken by B speakers. Each sample has labels for the class that it belongs to and the speaker whom it is from. The objective function is given by direct derivation from (4) as

$$L_1(\Lambda) = \sum_{b=1}^{B} \int \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \ell(\mathbf{O}_{b,k,i}, C_k, F(\Lambda,\Gamma)) p(\Gamma|\Lambda,V_b) d\Gamma, \quad (6)$$

where $n_{b,k}$ is the number of samples in class C_k, which is extracted from utterances spoken by speaker V_b, $\mathbf{O}_{b,k,i}$ denotes the ith feature vector in the samples, and ℓ is the loss function for individual samples. The distribution $p(\Gamma|\Lambda,V_b)$ here is generally unknown. Taking into account the simultaneous training of the distribution $p(\Gamma|\Lambda,V_b)$, sample sets $\mathcal{A}_{b,h}, h = 1, ..., H_b$ are prepared as the adaptation sample sets extracted for each input speaker $V_b, b = 1, ..., B$. Then the objective function is given as

$$L_2(\Lambda) = \sum_{b=1}^{B} \sum_{h=1}^{H_b} \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \ell(\mathbf{O}_{b,k,i}, C_k, F(\Lambda,\hat{\Gamma}_{b,h})), \quad (7)$$

where

$$\hat{\Gamma}_{b,h} = \operatorname*{argmin}_{\Gamma} \sum_{k=1}^{K} \sum_{i=1}^{m_{b,k}^{(h)}} \ell_a(\mathbf{O}_{b,k,i}^{(\mathcal{A}_{b,h})}, C_k, F(\Lambda,\Gamma)). \quad (8)$$

$\mathbf{O}_{b,k,i}^{(\mathcal{A}_{b,h})}$ is the ith feature vector in class C_k in the adaptation training sample set $\mathcal{A}_{b,h}$.

In addition, when the size of the given training sample set for the adaptation is large enough with regard to the degrees of freedom in Γ, the distribution $p(\Gamma|\Lambda,V_b)$ becomes very sharp. In this case, we assume that the distribution $p(\Gamma|\Lambda,V_b)$ is approximately a delta function, where $p(\Gamma|\Lambda,V_b) \neq 0$ at $\Gamma = \hat{\Gamma}_b$. Based on this assumption, the value of $\hat{\Gamma}_b$ is determined by using all samples of speaker V_b in $\mathcal{X}$ as an adaptation training sample set. Then the objective function is given as

$$L_3(\Lambda) = \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \ell(\mathbf{O}_{b,k,i}, C_k, F(\Lambda,\hat{\Gamma}_b)), \quad (9)$$

where

$$\hat{\Gamma}_b = \underset{\Gamma}{\text{argmin}} \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \ell_a(\mathbf{O}_{b,k,i}, C_k, F(\Lambda, \Gamma)) \qquad (10)$$

Furthermore, when $\ell_a(\mathbf{O}, C_k, F(\Lambda, \Gamma)) = \ell(\mathbf{O}, C_k, F(\Lambda, \Gamma))$, by combining (9) and (10) we can obtain the objective function as

$$L_3'(\Lambda, \Gamma) = \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \ell(\mathbf{O}_{b,k,i}, C_k, F(\Lambda, \Gamma_b)), \qquad (11)$$

where $\Gamma = \{\Gamma_1, ..., \Gamma_B\}$.

We note that all objective functions defined here are empirical average costs and that the minimization is only for the classification cost incurred in classifying the training samples. Although generalization should also be considered, we have used these objective functions with large amounts of training data as the first step in confirming the validity of the proposed approach.

3.3. Optimization Method

For the objective functions given in subsection 3.2, a local or global optimum solution can be obtained by using iterative neighborhood search strategies, such as the gradient descent method. In addition, for the objective function L_3', if optimization processes with respect to Λ and Γ both exhibit global convergence, carrying out both optimization processes alternately in an iterative fashion would also exhibit global convergence. This property becomes effective particularly in cases where a global solution can be easily found in each process that is alternately carried out. In the experiments described in the following section, we utilized this property. Note that this convergence property does not necessarily hold for the objective functions L_2 and L_3.

4. EXPERIMENTS

To evaluate the proposed training method, we conducted simple experiments in five-class, fixed-dimensional Japanese vowel pattern recognition with speaker adaptation, using classifiers trained based on the objective function L_3'. Vowel tokens were extracted from 216 isolated words spoken by 40 speakers (20 males, 20 females) and were digitized (fs 12 KHz). The center fragment of each vowel was selected using a 32 msec Hamming window and converted into a feature vector consisting of 16 order LPC cepstral coefficients. The number of tokens in the whole token set was 26,667. The classifier consisted of normal distribution functions with diagonal covariances as discriminant functions. The classifier parameter set Λ is represented as

$$\Lambda = \{\mu_1, \sigma_1, ..., \mu_K, \sigma_K\}, \qquad (12)$$

where μ_k and σ_k respectively denote the mean vector and the diagonal covariance vector in the discriminant function for class C_k.

$$\mu_k = [\mu_{k,1}, ..., \mu_{k,J}]^{\mathrm{T}}, \quad \sigma_k = [\sigma_{k,1}^2, ..., \sigma_{k,J}^2]^{\mathrm{T}} \qquad (13)$$

where $J(= 16)$ is the number of components in the feature vector. The superscript $^{\mathrm{T}}$ denotes the matrix transposition. The classifier adaptation to the speaker represented by V_b is carried out by moving all mean vectors through the addition of the single vector $\Gamma_b = [\gamma_{b,1}, ..., \gamma_{b,J}]^{\mathrm{T}}$: parameters of the adaptation transformation. ML and MCE were used as training criteria in the experiments. In the case of ML, (11) is written as

$$L_3'(\Lambda, \Gamma) = \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{i=1}^{n_{b,k}} \Big[\frac{J}{2} \log(2\pi) + \frac{1}{2} \log(|\Sigma_k|) + \frac{1}{2} (\mathbf{O}_{b,k,i} - \mu_k - \Gamma_b)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{O}_{b,k,i} - \mu_k - \Gamma_b) \Big], \qquad (14)$$

where $\Sigma_k = \mathrm{diag}(\sigma_{k,1}^2, ..., \sigma_{k,J}^2)$. In the case of MCE, the objective function is decided by using the smoothed loss function[7] for the gradient algorithm, which approximately represents the classification error.

For the optimization of the ML criterion, optimization processes with respect to both Λ and $\Gamma(= \{\Gamma_1, ..., \Gamma_B\})$ were carried out alternately in an iterative fashion. In each process, the solutions can be obtained using simple equations: Γ in particular was calculated with the following equation, which can be obtained by setting $\frac{\partial L_3'}{\partial \gamma_{b,j}} = 0$:

$$\gamma_{b,j} = \frac{\sum_{k=1}^{K} \frac{1}{\sigma_{k,j}^2} \sum_{i=1}^{n_{b,k}} (\mathbf{o}_{b,k,i,j} - \mu_{k,j})}{\sum_{k=1}^{K} \frac{n_{b,k}}{\sigma_{k,j}^2}}, \qquad b = 1, ..., B \quad , \quad j = 1, ..., J \qquad (15)$$

where $\mathbf{o}_{b,k,i,j}$ represents the jth component in the feature vector $\mathbf{O}_{b,k,i}$.

In the MCE criterion, the objective function was minimized by the gradient method with respect to all parameters using the following adjustment rule:

$$\{\Lambda, \Gamma\}_{t+1} = \{\Lambda, \Gamma\}_t - \epsilon \nabla L_3'(\Lambda, \Gamma), \qquad (16)$$

where $\{\Lambda, \Gamma\}_t$ denotes the parameter set at the tth iteration. The parameters were adjusted after the entire training set $\mathcal{X}$ was classified. The parameter values obtained in the ML training were used as the initial values.

As the solution is infinity for this optimization problem, after all processes were completed, the adaptation vectors $\Gamma_b, b = 1, ..., B$ were modified by adding vector $\mathbf{r}$ such that $\sum_{b=1}^{B} \Gamma_b = \mathbf{0}$. Inversely, the mean vectors in the classifiers were modified by deleting the same vector $\mathbf{r}$.

In the experiments, for comparison with the classifiers (CP) obtained by the proposed training methods, speaker inde-

pendent classifiers (CC) were obtained by the conventional training method based on (2).

The performance of each classifier was evaluated by *leave-one-out* testing: each classifier was trained using data from thirty-nine speakers ($B = 39$), adaptation was carried out to the other speaker V_b, and error rate was calculated for open data of speaker V_b. Thirty-two feature vectors for each class, randomly selected from the sample set obtained from speaker V_b, were used as adaptation training data to determine the value of the parameters of the transformation to adapt to speaker V_b. The adaptation vector Γ_b was decided by the same criterion as that by which the classifier had been trained. The average values of the error rates obtained in forty trials by changing the adaptation target speaker are shown in Figure 2.

In Figure 2, the **Adapted CC** and **Adapted CP** represent the CC and CP on which adaptation was carried out. We can see that the error rates obtained with the **Adapted CP** were lower than those obtained with the **Adapted CC**. The decrease in the value of the error rate from the **CC** to the **Adapted CP** is, compared with that from the **CC** to the **Adapted CC**, 74% larger in the ML case and 54% larger in the MCE case. These results confirm the validity of the proposed method for designing classifiers. We can also see that, as expected, the **CC** achieved better performance than did the **CP**. This is understandable because the CC was trained so as to get high performance without adaptation, whereas the CP was trained so as to get high performance with adaptation.

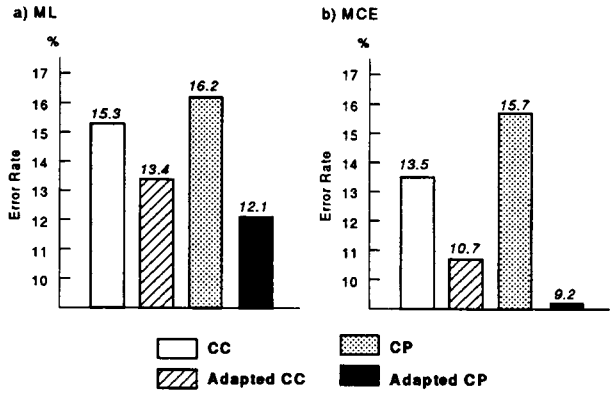

Figure 2: Average error rate in experiments with forty adaptation target speakers.

In addition, the respective mean and standard deviation of the difference between the error rates of the **Adapted CC** and the **Adapted CP** for all adaptation target speakers were 1.25% and 1.71% in the ML training cases, and 1.52% and 1.96% in the MCE training cases. These values show that the **Adapted CP** achieved better performance than the **Adapted CC** without large variance by changing the adaptation target speaker.

5. CONCLUSION

This paper proposed a method for designing a pattern classifier that achieves high performance after off-line transformation-based adaptation was carried out on it. Experimental results in a fixed dimensional vowel pattern recognition task with speaker adaptation clearly demonstrated its validity.

One can apply the proposed method to dynamic (variable-durational) patterns for practical speech recognition tasks using the hidden Markov model. It has been reported that changing the degrees of freedom of the trained transformation according to the size of adaptation training data leads to better performance (for example [8]). For such methods, it might be effective to prepare multiple classifiers trained according to the size of adaptation training data by the proposed method.

The method presented in this paper is quite general, and can be widely applied. This paper has described the case in which classifier parameters are adapted, but the presented method is also applicable to the case in which the feature extractor is adapted.

References

[1] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition*, Kuluwer Academic Publishers, 1996.

[2] K.-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. Signal Processing*, Vol.38, No.4, pp599-609, 1990.

[3] C.H. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol.9, No.2, pp.171-185, 1995.

[4] T. Matsui and S. Furui, "A study on speaker adaptation based on minimum classification error training," *Proc. EUROSPEECH'95*, 1995.

[5] S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *Proc. ICASSP*, pp.286-289, 1989.

[6] S.J. Raudy and A.K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.13, No.3, pp.252-264, 1991.

[7] B.-H. Juang and S. Katagiri, " Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing*, Vol.40, No.12, pp.3043-3054, 1992.

[8] H. Hattori and S. Sagayama, "Vector field smoothing principle for speaker adaptation," *Proc. ICASSP*, Vol.1, pp.381-384, 1992.