Manifold HLDA and its application to robust speech recognition

Toshiaki Kubo, Tetsuji Ogawa, Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan

Abstract

A manifold heteroscedastic linear discriminant analysis (MHLDA) which removes environmental information explicitly from the useful information for discrimination is proposed. Usually, a feature parameter used in pattern recognition involves categorical information and also environmental information. A well-known HLDA tries to extract useful information (UI) to represent categorical information from the feature parameter. However, environmental information is still remained in the UI parameters extracted by HLDA, and it causes slight degradation in performance. This is because HLDA does not handle the environmental information explicitly. The proposed MHLDA also tries to extract UI like HLDA, but it handles environmental information explicitly. This handling makes MHLDA-based UI parameter less influenced of environment. However, as compensation, in MHLDA, the categorical information is little bit destroyed. In this paper, we try to combine HLDA-based UI and MHLDA-based UI for pattern recognition, and draw benefit of both parameters. Experimental results show the effectiveness of this combining method.
Index Terms: HLDA, MHLDA, robust speech recognition.

1. Introduction

A statistical pattern recognition has a problem that the discriminative performance becomes worse when the training environment differs from the evaluation environment.

In order to solve this problem, adaptation techniques such as MLLR [1] are widely applied. However, it is not realistic to constantly obtain the adaptation data in the situation when the environment is changing at any time. On the other hand, the framework of HLDA [2] can extract useful information (UI), which represents categorical information of each class, from a feature parameter. In this framework, nuisance information (NI), which represents environmental information common to all classes, can be removed implicitly. Therefore, HLDA can give the high performance robustly to the changes of the environments without evaluation environment information. Here, we focus on the information on the known environmental attributes such as the room acoustics, the characteristics of speakers and the background noises in speech recognition. We call such information a known environmental information (KEI). In HLDA, KEI is due to be regarded as NI, however, it can not always be precisely removed. If the KEI is unfortunately remained in the UI parameters, the discriminative performance is possible to be degraded.

In order to solve this problem, we proposed a manifold HLDA (MHLDA) which does not require the constraint of HLDA that the information of the feature parameters is divided into only 2 classes; the UI and the NI. In this paper, using the MHLDA proposed here, the information of the feature parameters is divided into 3 classes; the UI, the KEI and the NI. Such generalization of HLDA follows that not only the UI but also the KEI can be extracted precisely, and therefore the KEI can be removed explicitly from a feature parameter unlike the HLDA which removes the KEI implicitly from a feature parameter. Furthermore, an integration of HLDA and MHLDA is performed aiming at reducing the errors given by each method in a complementary style.

This paper is organized as follows. As the base of the proposal in this paper, useful information extraction and HLDA is briefly surveyed in section 2. In section 3, the manifold HLDA is proposed. The motivation for generalization and the formulation are described in this section. Section 4 gives the results of the proposed method using a spoken word recognition. Finally, in section 5, concluding remarks are presented.

2. Useful information extraction

In this paper, the collective term of the frameworks which can extract useful information from a feature vector is defined as a useful information extraction (UIE). A well-known HLDA can be regarded as one of the UIEs. In this section, as the basis of the proposed method, the overview of HLDA is described.

HLDA can divide the factors of a feature vector into 2 classes by the coordinate transformation of the feature vector. One class has the UI parameters which contribute to the discrimination and the other class has the NI parameters which does not contribute to the discrimination. Then, only the UI parameters extracted by HLDA are used for the feature parameters. Therefore, HLDA can realize the pattern recognition framework robust to the environmental changes

since the useful information includes only categorical information but not include the environmental information, ideally.

HLDA is the framework in which a maximum likelihood linear transform (MLLT) is applied to LDA and HDA [3]. Also, in HLDA, an equal class variance constraint assumed by LDA can be removed.

In HLDA, the factors of a n dimensional feature vector can be divided into p dimensional UI parameters and $n-p$ dimensional NI parameters, where $p < n$. A transformation matrix θ can be represented as follows:

$$\theta = [\theta_1 \dots \theta_n] = [\theta_p \theta_{n-p}]$$

where θ is a $n \times n$ matrix, θ_p is a $n \times p$ matrix and θ_{n-p} is a $n \times (n-p)$ matrix. In order to give above constraint, in the framework of HLDA, the p dimensional UI parameters are different for each class and $n-p$ dimensional NI parameters are common to all the classes in the means and the variances after the coordinate transformation. Here, the mean μ_j and the variance $\mathbf{\Sigma}_j$ of class j are defined as follows:

$$\mu_j = \begin{bmatrix} \mu_j^p \\ \mu_0^{n-p} \end{bmatrix}, \quad \mathbf{\Sigma}_j = \begin{bmatrix} \mathbf{\Sigma}_{j(p\times p)}^p & 0 \\ 0 & \mathbf{\Sigma}_{0(n-p\times n-p)}^{n-p} \end{bmatrix}$$

where μ_0 and $\mathbf{\Sigma}_0$ is common to all the classes.

If $\mathbf{y}_i$ is obtained by transforming $\mathbf{x}_i$, the logarithmic likelihood of all data for each class in each environment, under the linear transformation and under the Gaussian model assumption, is defined as follows:

$$\begin{aligned} &L(\mu_j, \mathbf{\Sigma}_j, \theta | \mathbf{x}_i) \\ &= -\frac{1}{2} \sum_{i=1}^{N} \{ (\theta^T \mathbf{x}_i - \mu_{g(i)})^T \mathbf{\Sigma}_{g(i)}^{-1} (\theta^T \mathbf{x}_i - \mu_{g(i)}) \\ &\quad + \log((2\pi)^n |\mathbf{\Sigma}_{g(i)}|) \} + \log|\theta| \end{aligned} \tag{1}$$

where $\mathbf{x}_i$ attributes to the class $g(i)$ and N denotes the number of all data. In the parameter estimation based on a maximum likelihood criterion, equation (1) is defined as an objective function and then θ which maximizes this function is required to be obtained. By differentiating the likelihood function with respect to the parameters μ_j and $\mathbf{\Sigma}_j$, the means and the variances can be estimated. Substituting these estimates obtained above in equation (1) can give the estimate of θ as follows:

$$\begin{aligned} \hat{\theta} = \arg\max_{\theta} \Bigg\{ &- \frac{N}{2} \log|\theta_{n-p}^T \mathbf{T} \theta_{n-p}| \\ &- \sum_{j=1}^{J} \log|\theta_p^T \mathbf{W}_j \theta_p| + N \log|\theta| \Bigg\} \end{aligned} \tag{2}$$

where $\mathbf{W}_j$ denotes the variance of class j before coordinate transformation and $\mathbf{T}$ denotes the variance in all data before coordinate transformation. Here, quadratic algorithms are applied to the optimization of θ.

3. Manifold HLDA

3.1. The motivation of proposing MHLDA

To take a speech recognition as an example, a feature parameter has not only a phoneme information as a categorical information but also the various environmental information, which does not contribute to the discrimination, such as the characteristics of speakers, the room acoustics, the kinds or the loudness of background noises and the characteristics of microphones. We call such information a known environmental information (KEI). HLDA tries to extract useful information from a feature parameter. Therefore we expect that the KEI can be removed from the UI parameter implicitly. However, the KEI is still remained in the UI parameters, and it causes slight degradation in performance. So we attempt to extract not only the UI parameters but also the KEI parameters aiming at removing the KEI from the UI explicitly.

HLDA has a constraint that the factors of a feature parameter is divided into only 2 classes; the UI and the NI, and therefore can not realize above framework. So in this paper, we propose manifold HLDA (MHLDA) which removes the constraint of the HLDA. MHLDA can divide the information of a feature parameters into 3 classes; the UI, the KEI and the NI, and therefore can remove the KEI precisely from a feature vector.

3.2. Formulation

In MHLDA, a n dimensional feature vector is divided into p dimensional UI parameters, q dimensional KEI parameters and $n-p-q$ dimensional NI parameters, where $p+q < n$. A transformation matrix θ is described as follows:

$$\theta = [\theta_1 \dots \theta_n] = [\theta_p \theta_q \theta_{n-p-q}]$$

where θ is a $n \times n$ matrix, θ_p is a $n \times p$ matrix, θ_q is a $n \times q$ matrix and θ_{n-p-q} is a $n \times (n-p-q)$ matrix. Here, we give the 3 constraints in the means and the variances of each class in each environment after the coordinate transformation as follows. First, the p dimensional UI parameters are different for each discriminative class. Secondly, the q dimensional KEI parameters are different for each environmental attributes. Finally, the $n-p-q$ dimensional NI parameters are common to all the classes. Here, the mean $\mu_{j,k}$ and the variance $\mathbf{\Sigma}_{j,k}$ of class j in environment k are represented as follows:

$$\mu_{j,k} = \begin{bmatrix} \mu_j^p \\ \mu_k^q \\ \mu_0^{n-p-q} \end{bmatrix}$$

$$\mathbf{\Sigma}_{j,k} = \begin{bmatrix} \mathbf{\Sigma}_{j(p\times p)}^p & 0 & 0 \\ 0 & \mathbf{\Sigma}_{k(q\times q)}^q & 0 \\ 0 & 0 & \mathbf{\Sigma}_{0(n-p-q\times n-p-q)}^{n-p-q} \end{bmatrix}$$

where μ_0 and $\mathbf{\Sigma}_0$ are common to all the classes in all environment. When θ transforms $\mathbf{x}_i$ into $\mathbf{y}_i$, the logarithmic likelihood of all data under the linear transformation and under the Gaussian model assumption for each class in each environment is described as follows:

$$\begin{aligned} &L(\mu_{j,k}, \mathbf{\Sigma}_{j,k}, \theta|\mathbf{x}_i) \\ &\quad = -\frac{1}{2}\sum_{i=1}^{N}\{(\theta^T\mathbf{x}_i - \mu_{g(i)})^T\mathbf{\Sigma}_{g(i)}^{-1}(\theta^T\mathbf{x}_i - \mu_{g(i)}) \\ &\qquad + \log((2\pi)^n|\mathbf{\Sigma}_{g(i)}|)\} + \log|\theta| \end{aligned} \tag{3}$$

where N denotes the number of all data and $\mathbf{x}_i$ attributes to the class $g(i)$. $g(i)$ denotes the suffix expressing the product of a discriminative class and a environment. The following equation can be conclusively obtained by the optimization based on maximum likelihood criterion same as HLDA.

$$\begin{aligned} \hat{\theta} = \arg\max_{\theta} \Bigg\{ &-\frac{N}{2}\log|\theta_{n-p-q}^T\mathbf{T}\theta_{n-p-q}| \\ &-\sum_{j=1}^{J}\log|\theta_p^T\mathbf{W}_j\theta_p| - \sum_{k=1}^{K}\log|\theta_q^T\mathbf{W}_k\theta_q| \\ &+N\log|\theta|\Bigg\} \end{aligned} \tag{4}$$

where $\mathbf{W}_j$ is the variance of class j, $\mathbf{W_k}$ is the variance of environment k and $\mathbf{T}$ is the variance of all data before coordinate transformation. Equation (4) consists of following 4 terms. First term is related to NI. Second term is related to UI. Third term is related to KEI. And last term is the regularization term.

4. Spoken word recognition experiment

4.1. Overview of experiment

In order to evaluate the effectiveness of MHLDA and the integration of HLDA and MHLDA, the spoken word recognition experiment was performed. Here, we focused on the speaker information as the environmental factor degrading the classification performance. We examined whether the MHLDA can explicitly remove the speaker information from a feature parameter and therefore can precisely extract the categorical information (phoneme information in the speech recognition).

4.2. Speech data

The experimental comparisons are conducted using the ATR phoneme balanced words database. Here, we used 6960 words from above database. We picked up 20 speakers and 116 words, 3 times per speaker.

The acoustic feature parameters used for the training and test are represented by 39 dimensional parameters (12 MFCCs, power, 12 Δ MFCCs, Δ power, 12 $\Delta\Delta$ MFCCs and $\Delta\Delta$ power), sampled every 10 ms.

Table 1: *The evaluation items of spoken word recognition*

evaluation items	transformation	useful dimension
BASE	-	39
ADAPT	MLLR	39
UIE-HLDA	HLDA	36
UIE-MHLDA	MHLDA	33
UIE-INTEG	Integration of HLDA and MHLDA	69

4.3. Statistical acoustic model

We used 3 kinds of statistical acoustic models: a base model, an adapted model and a UIE model.

A statistical acoustic model is based on a monophone hidden Markov model (HMM) according to each phoneme using 39 dimensional feature parameters. Each phoneme model consists of 3 states. The distribution function in each state is represented by a 16-mixture Gaussian distribution with diagonal covariance. We call this model a base model. The base model are trained with 10742 newspaper article sentences from the ASJ database (ASJ-JNAS)[4].

The base model is adapted using a small amount of evaluation speaker data. We call this model an adapted model. In order to obtain this model, we used 20, 40, 60, 100 and 200 phoneme balanced words which were not used for the training of the statistic models and the evaluation.

The model which is trained using the UI parameters extracted by UIE described in Sect. 2 is called a UIE model. Training data for transformation matrix is same as the training data for base model. Here, the useful dimension is fixed to the value with which the best recognition performance is obtained.

4.4. Evaluation items

Evaluation items are shown in Table 1. BASE is the framework in which the classification is performed using the base model with 39 dimensional feature parameters. ADAPT is the framework in which the classification is performed using the adapted model. UIE-HLDA and UIE-MHLDA are the framework in which the training and classification are performed using the feature parameters transformed by the HLDA and MHLDA based transformation matrix. In this experiment, 36 and 33 dimensional useful information are used for HLDA and MHLDA respectively, since the best performance was obtained in the preliminary experiments. Also, in UIE-INTEG, a feature vector is obtained by simply combining the feature vector with the UI parameter extracted by HLDA and that extracted by MHLDA. The dia-

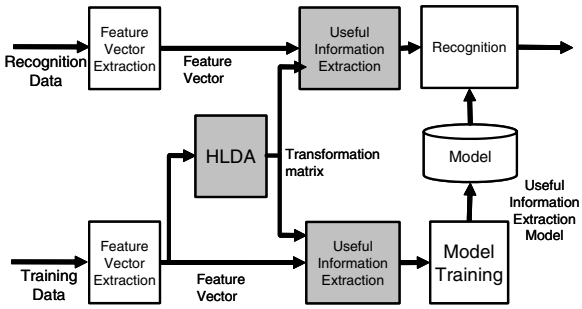

Figure 1: *The diagram of the experiments using UIE frameworks*

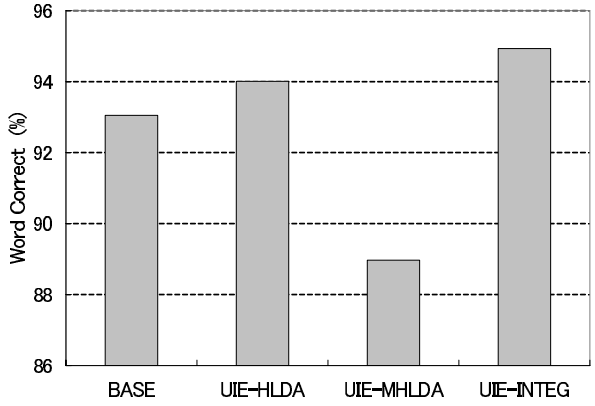

Figure 2: *Word recognition rate for various frameworks*

gram of the experiments are shown in Fig.1. As we can see in the diagram, the framework of UIE is not required the data of evaluation speaker.

4.5. Experimental result

Figure 2 shows the word recognition rate for various frameworks. Also, Fig. 3 shows the word recognition rate as a function of the number of adaptation data. From Fig.2, HLDA (94.0%) reduced the errors compared with BASE (93.1%) while MHLDA (89.0%) degraded the performance compared with BASE. However, the integration of HLDA and MHLDA (UIE-INTEG) achieved the good performance of 94.9%. This is 27% reduction in error rate compared with that of BASE without evaluated speaker information. Here, ADAPT gives the best performance (95.5%). However, to obtain the performance, 200 words are required for adaptation. It is not realistic to obtain such a large amount of data for every speaker. Also, as we can see in these figures, it was incidentally found that the UIE-INTEG exceeds the performance obtained when 100 words are used for adaptation (94.6%). As a result, integrating 2 kinds of UIE gives the performance as well or better than that using the practical

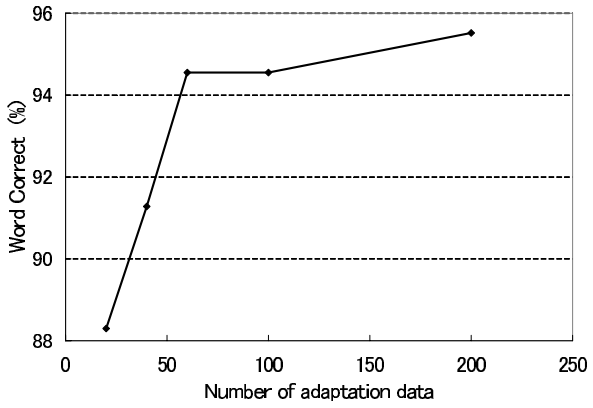

Figure 3: *Word recognition rate as a function of the number of adaptation data*

speaker adaptation technique without any evaluated speaker information.

5. Conclusion

In this paper, we proposed a manifold HLDA (MHLDA) aiming at removing environmental information explicitly from the useful information parameters for discrimination. Furthermore, we tried to combine HLDA-based UI parameters and MHLDA-based UI parameters.

The proposed method was applied to the spoken word recognition. The results showed that the integration of the 2 kinds of UIE, HLDA and MHLDA, reduced errors compared with each UIE method, and gave the robust performance to the variations of speakers.

6. References

[1] M.J.F.Gales, P.C.Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech and Language, vol.10, pp.249-264, 1996.

[2] N.Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, Johns Hopkins Univ., 1997.

[3] G.Saon, M.Padmanabhan, R.A.Gopinath, S.Chen, "Maximum likelihood discriminant feature spaces," Proc. ICASSP, vol.2, pp.1129-1132, 2000.

[4] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," Proc. ICSLP, pp.3261-3264, 1998.